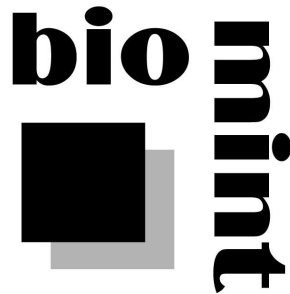


**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---



**Deliverable Report 10.6 (Year 1)**  
**Yearly patent report**

**Authors:**       **Compiled on behalf of the consortium by**  
                          **Kristof Van Belleghem**  
                          **PharmaDM**

**Status:**         **Final**

**Distribution:** **Consortium**

**Version:**       **1.0**

**Checkers:**

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

**PROJECT MANAGER**

*Name: Professor Terri Attwood*

*Address: Bioinformatics Group, School of Biological Sciences,*

*Stopford Building University of Manchester, Oxford Road, Manchester, M13 9PT*

*Phone Number: +44 161 275 5082*

*Fax Number: +44 161 275 5082*

*E-mail: attwood@bioinf.man.ac.uk*

**TABLE OF CONTENTS**

<u>1. Executive Overview</u> . . . . .	3
<u>2. Relevant Patents</u> . . . . .	3
<u>2.1 US2003066025</u> . . . . .	3
<u>2.2 US2003216905</u> . . . . .	3
<u>2.3 US2003014398</u> . . . . .	4
<u>2.4 US6553385</u> . . . . .	4
<u>3. Semi-relevant Patents</u> . . . . .	5
<u>3.1 EP0996899</u> . . . . .	5
<u>3.2 US6442545</u> . . . . .	5
<u>3.3 US6601026</u> . . . . .	6
<u>3.4 WO03012661</u> . . . . .	6
<u>3.5 US2002156817</u> . . . . .	6
<u>3.6 US2002143524</u> . . . . .	7
<u>3.7 US2002010574</u> . . . . .	7
<u>3.8 US6026388</u> . . . . .	7
<u>3.9 US2003093276</u> . . . . .	8
<u>4. Conclusion</u> . . . . .	8

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

## **1. Executive Overview**

In order to stay well-informed about the commercialization potential of the BioMinT tool, we checked the international patent databases for patents showing some relation to (parts of) the current and planned tool. For the 2003 report, we thoroughly checked 2003 patents and, since some time has passed between first submission and start of the project, we also looked (somewhat less thoroughly) through patents of the period 2000-2002. Eventually we found 4 patents we considered relevant and 9 semi-relevant ones. We briefly discuss the important correspondences to and differences from the BioMinT tool below.

## **2. Relevant Patents**

### **2.1 US2003066025**

- Method and system for information retrieval
- published 2003-04-03

This patent describes a system for query expansion. One part of the system expands a natural language query to include terms and concepts related to keywords in the query (e.g. using synonyms found in various databases). In the BioMinT tool workflow, this is a first step preceding document retrieval. The biomedical literature is explicitly mentioned in this patent as an application area of choice.

### **2.2 US2003216905**

- Applying a structured language model to information extraction
- published 2003-11-20

This patent describes a system for information extraction, based on parsing free text using a structured language model. The model is obtained by combining syntactic annotations with semantically annotated training data, an approach currently suggested as a basis for our natural language processing component. However, this combination of two types of annotation is the only overlap with the BioMinT tool, as the described system is geared toward understanding and reacting to singular, command-like sentences, as used in a user interface. This task is quite different from extracting information from huge volumes of literature, where

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

- sentences are not guaranteed to be relevant
- sentences are much more complex than simple commands
- sentences typically do not map to a single semantic frame
- high volumes of text must be processed in an acceptable time.

### **2.3 US2003014398**

- Query modification system for information retrieval
- published 2003-01-16
- also published as : JP2003016089

This patent describes a query processing system transforming a query into a keyword vector, and an interface for the user to attach weights to the various keywords in the vector. This vector can then be used to search PubMed, for example. There is a limited resemblance to a step in the BioMinT query expansion, where the user can select or deselect the automatically proposed synonyms and related terms, before submitting the query. However, the described system does not itself suggest expansion terms for the query.

### **2.4 US6553385**

- Architecture of a framework for information extraction from natural language documents
- published 2003-04-22
- also published as : US2002007358

The described framework addresses some of the same issues as the BioMinT tool architecture developed in WP2, in particular by providing two APIs: one “outside” for application programs to use the framework, and one “inside” allowing various information extraction modules to be plugged in. The framework handles control of the various extractors.

However, the system is only a framework handling the information extraction part of a text mining task, not a full-fledged application like the BioMinT tool. The framework fails to handle query expansion or information retrieval at all, nor does it have provisions for calling other external applications. Moreover, it does not deal with the actual database annotation tasks (it is in fact not supposed to be biomedicine-specific), and has no global background knowledge module.

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

### **3. Semi-relevant Patents**

#### **3.1 EP0996899**

- Apparatus and methods for an information retrieval system that employs natural language processing of search results to improve overall precision
- published 2000-05-03
- also published as : WO9905618, US5933822

The described system uses a deep NLP analysis to match documents with a natural language query. These documents should be obtained by a keyword-based information retrieval step, which is not included in the system. As foreseen in the BioMinT tool as well, documents will be scored and ranked based on the number of corresponding semantic structures in query and documents. This is roughly comparable to our information extraction step, although the BioMinT tool will not rely on parsing the query, will only apply shallow parsing, and will use biomedical background knowledge to enhance retrieval and extraction.

#### **3.2 US6442545**

- Term-level text with mining with taxonomies
- published 2002-08-27
- note : applicant is Clearforest

The described system mines texts for information on one or more terms of a taxonomy, in order to discover relationships between these terms and other terms from the taxonomy. Potentially these relationships can be used to extend the taxonomy. The system is mainly based on co-occurrence of terms, although it is claimed to use grammatical structure of sentences, in particular in order to extract sets of document labeling terms.

The use of taxonomies shows some similarity with an envisioned BioMinT approach to information extraction using background knowledge (in particular, more specific terms are also looked for when a search term is specified), but is not really geared toward biomedicine and its specific information extraction needs. Moreover, the system does not use grammatical analysis for information extraction, but only for the very limited task of document labeling. Finally, observe that the system handles no information retrieval either.

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

### **3.3 US6601026**

- Information retrieval by natural language querying
- published 2003-07-29
- also published as WO0120500, US2003078766

The described system is a general-purpose query answering system, answering natural language queries based on information in a database of text documents. New documents can be acquired from a web crawler or news service. An index is maintained capturing all relevant information of documents added to the system; this information is extracted using some kind of grammar. The query is typically analysed using a similar grammar, and matched against the index. Various types of output can be generated, like a summary for a group of documents, quotations from various documents, or full documents with relevant portions highlighted.

The information extraction part of this system is an alternative which has been considered for BioMinT, but is for now not pursued.

### **3.4 WO03012661**

- Computer based summarization of natural language documents
- published 2003-02-13

This system builds document summary reports by performing deep linguistic analysis on the text, weighting each sentence (with weights dependent on the desired summary type), and generating a report using the most important sentences. Analysis uses domain knowledge in some form (i.e. “is based on an understanding at the level of objects, facts, and regularities of the knowledge domain to which the document refers”).

A variant of this approach could be an option for implementing PRECIS report generation.

### **3.5 US2002156817**

- System and method for extracting information
- published 2002-10-24

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

The system generates database records from unstructured or semi-structured text files, in particular (but not limited to) e-mail messages. It does this by looking at context (e.g. category of the text) and using limited language analysis. However, the system's preferred embodiment specialises in short, telegraphic messages, not large volumes of text.

### **3.6 US2002143524**

- Method and resulting system for integrating a query reformation module onto an information retrieval system
- published 2002-10-03

This system transforms natural language queries into expanded keyword-based queries, by performing syntax analysis on the input query, determining important terms, expanding those terms using synonym and typing information, and filtering unimportant terms. An expanded boolean query is then constructed which should better reflect the desired results. This system is comparable to the query expansion step in the BioMinT tool, although the latter does not start from a natural language query.

### **3.7 US2002010574**

- Natural language processing and query driven information retrieval
- published 2002-01-24
- also published as WO0182123

This patent describes a system for converting natural language sentences into structured frames containing subject, action, object and up to 5 other fields. The resulting frames are stored in a database, which can then be queried; the queries can be similar natural language requests, which are converted like the data sentences. Again, there is a similarity between this approach and one option considered for using the NLP module (i.e. precomputing all relevant information), although it does not take background knowledge into account.

### **3.8 US6026388**

- User interface and other enhancements for natural language information retrieval system and method

**Deliverables Report**  
**QLRI-2002-02770 BioMinT**  
**February 2004**

---

- published 2002-02-15

This is an interactive query enhancement system, analysing documents and user queries at the morphological, lexical, syntactic, semantic, discourse, and pragmatic levels. Queries are transformed into an enriched, structured format, which can then still be modified by the user. It is also possible to launch a query and then select the most relevant documents among those returned. The representation of these documents (obtained by a similar transformation) can then be integrated in the query to form a new query. The principle of this system (interactive query tuning) is also the one for information retrieval in the BioMinT tool.

### **3.9 US2003093276**

- System and method for automated answering of natural language questions and queries
- published 2003-05-15

This is another system for dealing with natural language queries. The idea here is to transform the query into a generic answer form, i.e. the expected form an answer to this query would take, through grammatical transformations on the query. Then, one can search documents for sentences matching that form (keywords search followed by checking if the form is matched). Relevant answer phrases, their sentences, and links to the original documents are returned.

## **4. Conclusion**

The above patents describe systems showing most similarity to the BioMinT tool, parts of the tool, or possible variant approaches considered in the tool's context. When commercialising the tool, these patents should be kept in mind in order to avoid conflicts. Meanwhile, the patents may suggest variants for some of the decisions made in the past or yet to be made.

This patent watch will be continued and reported on in the next two years of the project.