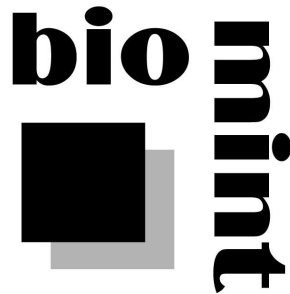


Deliverables Report
QLRI-2002-02770 BioMinT
February 2004



Deliverable Report 8.2
Requirements specification proposed by EUC members

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

1. <u>Executive Overview</u>	3
2. <u>Users and uses</u>	3
3. <u>Information Sources and Desired Answers</u>	3
4. <u>User Interaction</u>	4
5. <u>Platform/Performance requirements</u>	5
6. <u>Conclusion</u>	5

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

1. Executive Overview

In order to be able to increase the value of the BioMinT tool for the life science community in general, we have investigated and collected the expectations of potential users. For the largest part we organized this by asking end user club members to fill out a requirements questionnaire. This questionnaire featured questions concerning the expected users and uses in the respondent's company, the particular text sources to be mined, the types of documents and information to be retrieved resp. extracted, and the desired interaction with the system and presentation of results. Most responses came from employees of pharmaceutical and biotechnology companies, in roughly equal numbers. Below we summarize the answers.

2. Users and uses

As is to be expected, the user base at end user club companies is more diverse than at partners in the project, where database curators are by far the most important target group. End user club members indicate an approximately equal interest from curators, patent experts, bench researchers, literature researchers, information experts and bioinformaticians, and even an occasional interest from sales & marketing staff.

The expectations people have of a text mining tool vary accordingly, and in fact seems mostly inspired by the possibilities of the tools they currently use. For example, current Omniviz users stress text clustering/categorization, whereas frequent PubMed users stress ontology- and synonym-based information retrieval and extraction, and in general an improved PubMed-like system.

3. Information Sources and Desired Answers

The most important information sources to be mined according to users are PubMed and Chemical Abstracts. Other suggested sources, like BIOSIS, Biological abstracts and patent databases (to name just a few) are less critical, but desirable. In fact, ideally users would like to have no limits at all, i.e. they also mention interest in news sites and even "the web in general". Moreover, various users stress the importance of mining not only abstracts, but - whenever they are available - also full texts.

In general, when a text mining tool presents some answers, users find it of vital importance to have access to the supporting evidence, at least through links to all sources the answer is

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

based on, so they can check the original documents. Where possible, they also want available cross-references to other public databases to be presented. In other words, it is essential that all results of the tool can be checked by humans for their precise meaning and potentially related and additional information.

Some users would also like a connection to other bioinformatics tools like Blast/Fasta, sequence analysis and clustering tools, although others admit this may be beyond the scope of a text mining project.

Concerning the types of information to be retrieved or extracted, everyone stresses these types can not simply be fixed once and for all, and should therefore be easily extendable as the tool evolves. Likewise, no one can precisely define satisfactory levels of recall and precision of information retrieval and extraction. In any case, a trade-off of precision vs recall should be possible dependent on the amount of data. A good way to approach this would be via a scoring system marking the confidence of the system that a certain document or information item is relevant. A score threshold could then be used to determine the returned answer set.

4. User Interaction

Presentation of results should come in various forms and levels of detail:

- a graphical overview of the answer space which can be easily navigated (preferably already during query refinement at a high level)
- a ranking of documents with access to the documents and highlighting of the relevant sections
- english natural language-like statements and formatted entries ready for database input.
- graphs and statistics

It should also be possible to save the results in a format that can be shared with other users, and navigated as before when re-opened.

Users want to be able to exert some control over the text mining process, in particular by specifying weights on precision vs recall, and by selecting the available sources to be mined.

The use of background knowledge tuned to the wishes of the user to guide information extraction and retrieval is also considered very important: users wish to be able to import their own dictionaries and ontologies. Since it is a major effort to build these from scratch, these knowledge sources are preferably adapted/extracted from existing public sources. To generate such user-specific ontologies, an ontology editor would be a welcome addition to the tool.

Deliverables Report
QLRI-2002-02770 BioMinT
February 2004

5. Platform/Performance requirements

A vast majority of end users want to use the tool on an MS Windows platform, and are not really interested in other platforms. As for a preference for a stand-alone, client/server or web application, opinions diverge widely, and all variants seem to have their use in certain circumstances.

Users estimate that the number of abstracts to be mined in response to one query could be up to thousands, sometimes hundreds of thousands of abstracts, so the system must be able to cope with these amounts. However, users claim to be willing to wait for an answer, and definitely don't want to trade quality/quantity of results for a faster reply. On the other hand, some would be willing to accept lower answer quality if this could entail much less work for the user.

6. Conclusion

Apparently requirements of potential users are largely inspired by the possibilities of text mining tools they use or know of. Since current text mining systems come in many different flavours and do many different things, users expect a similarly broad range of functions from "the ideal text mining tool". The main strength of the BioMinT tool as it is planned now, is presumably the quite precise extraction of nuggets of information in response to queries that go beyond mere keyword combinations, and the use several sources of background knowledge. Hence, where choices of functionality need to be made, it seems most wise to concentrate on information-rich tasks.

Concerning many other parameters, like sources to be searched, background knowledge to be used, and types of information to be retrieved/extracted, the main message seems to be we need to make the system as easily extendable as possible, trying to restrict the impact of any particular choices on the overall system wherever we can.

The various other requirements, in particular those concerning user interaction with the system, will need to be prioritised according to importance and feasibility. Discussions on this issue among the partners will be initiated.