

**Welcome** to the second edition of the NaCTeM Newsletter. In this issue we focus on some of the broader themes surrounding the area of metadata extraction in the context of a number of practical projects at NaCTeM drawing upon associated research activity. This ranges from ideas surrounding alternative way to extract key terms and phrase through to the more semantic issues of extracting specific information or the relationships between entities. Further to this we discuss how linking text mining tools with external resources such as dictionaries or ontologies can benefit the results of tools and further support the use of text mining in annotation of large scale document collections.

More information about the centre can be found on our website at <http://www.nactem.ac.uk> or via a selection of surveys and briefing papers available at:

- Text Mining Briefing Paper - <http://jisc.ac.uk/media/documents/publications/bptextminingv2.pdf>
- Introduction to NaCTeM - <http://jisc.ac.uk/media/documents/publications/bpnationalcentrefortextminingv1.pdf>
- Vision for the future - <http://www.ariadne.ac.uk/issue53/ananiadou/>

## CheTA – Chemistry using Text Annotations

CheTA brings together a consortium of the Unilever Centre for Molecular Science Informatics, University of Cambridge (leading) and NaCTeM in close partnership with the Royal Society for Chemistry and Thomson Reuters. The CheTA project will integrate Cambridge's chemical text mining tool OSCAR with the U-Compare (<http://u-compare.org/>) workflow infrastructure developed by NaCTeM and others. This integration adds customised chemistry functionality to the world's largest public collection of interoperable text mining tools and promises to be of great value to stakeholders both in the JISC community and the wider chemistry and biomedical communities.

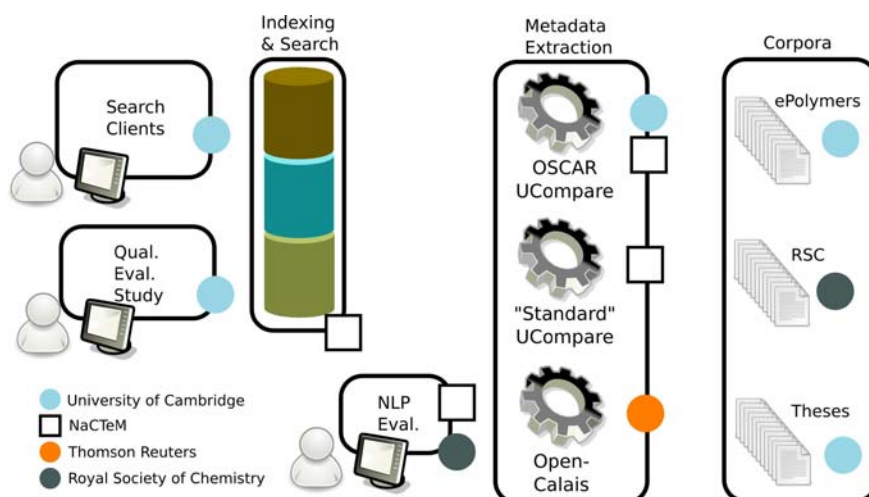


Figure 1: Architecture of CheTA project components

Following a baseline study and once the integration has been accomplished, the project will use the CheTA tools to index a corpus of documents of different types and provenance. CheTA will develop a rigorous evaluation framework with annotation studies for a formal scientific evaluation of the system, user requirements studies for the metadata needs of “real world users” and for comparing extracted metadata against the usefulness. Furthermore, the CheTA system will be compared with the performance of the Thomson Reuters OpenCalais service (<http://opencalais.com>) enhanced with a chemistry lexicon. Finally, we will quantify the economic cost of metadata generation by both human indexers and robots. We expect that the application of professionally maintained, automated and sustainable text mining services, enabled by CheTA to public information sources such as PubMed, will lead to significant future enhancements in resource discovery.

## Academic Promotion for Centre's Director

Our congratulations go to Sophia Ananiadou, director of NaCTeM, on the recent news of her academic promotion. Sophia has been made a Professor in the School of Computer Science at the University of Manchester in recognition of her extensive contributions to the school, the university and the discipline of text mining. We wish Sophia the very best wishes in her new role and look forward to the exciting future with her continuing to lead NaCTeM.



## Next Generation of Tag Clouds?

Tag clouds are becoming an increasing popular way of summarising the main focus or interests of a selection of online content. Algorithms generally use frequency of occurrence as the base form of importance of words or short phrase to the text. Whilst this can generate useful results it is possible that they can misrepresent the content by breaking up multiword phrases e.g. 'text mining' becomes 'text' and 'mining', potentially representing very different concepts. More advanced tools are beginning to take this into account and are returning frequencies of occurrence for phrases; however, NaCTeM's TerMine tool takes this a step further.

With TerMine the emphasis still remains with phrases, or as we call them 'multiword terms', but instead of using frequency of occurrence to mark the significance of a term we use an algorithm that takes into account nested terms to predict the overall significance. This has been evaluated for use across a number of projects and has been adopted by all of the projects in the A1 strand of the recent JISC call for automated metadata extraction. Visit the NaCTeM website to learn more about TerMine and try it for yourself at <http://www.nactem.ac.uk/software/terminer>.

Recently we have carried out a number of tests using the output of TerMine to drive tag clouds of famous speeches, blogs and twitter archives. As part of this work we have examined the use of the Wordle library (<http://www.wordle.net>) from IBM which creates a visually attractive layout for the tag clouds. Figure 1 shows the results of this process based upon a sample of blog posts from the blog of BBSRC Chief Executive, Professor Kell.

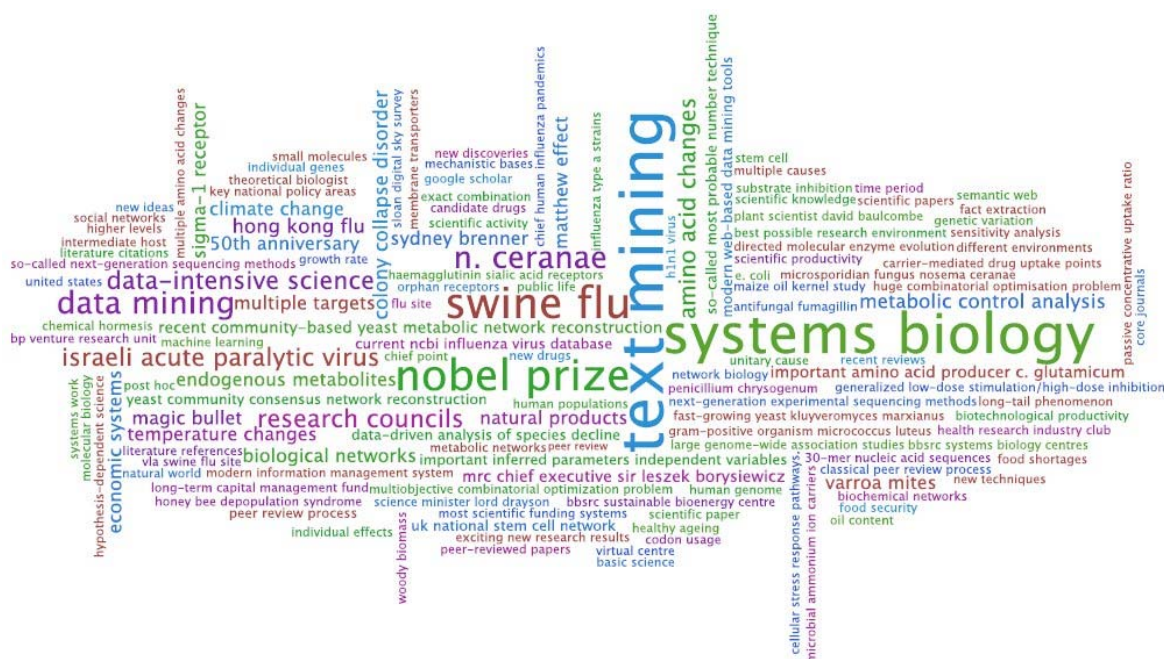


Figure 2: Wordle tag clouds using TerMine terms



## New Addition to the NaCTeM Team

We extend a warm welcome to the latest recruit at NaCTeM, Sasirekha Palaniswamy, who joins us in the role of Biomedical Text Mining Facilitator. The post has been jointly funded through our strategic partnership with the Biomedical Research Centre, and we anticipate considerable benefits to both communities through this collaboration. Sasirekha is responsible for the further support and extension of our widespread communities in the biomedical and health sciences. She would like to hear from those wishing to learn more about text mining or how it could benefit everyday research activities.

Email: [Sasirekha.Palaniswamy@manchester.ac.uk](mailto:Sasirekha.Palaniswamy@manchester.ac.uk)

## Automated Entity and Relationship Extraction for Biomedicine

Research in Systems Biology requires the use and integration of information in a wide range of structured data sets and databases as well as in the rapidly expanding, but less structured biomedical literature. The ONDEX SABR project aims to support this integration through an e-tool, ONDEX, which provides:

- graph-based database integration and visualization,
- workflows using Taverna,
- graph-based and statistical analysis of biological networks, and
- text mining.

The ONDEX SABR project brings together teams of technical and biological experts from three institutions: University of Manchester, Newcastle University, and Rothamsted Research. NaCTeM provides the text mining tools and expertise for this large project.

The NaCTeM text mining efforts in the ONDEX project aim to enable automatic recognition and extraction of a range of novel entities and relations associated with the core biological use cases within the project. These use cases include:

- automatic identification of organisms and associated habitats in collaboration with researchers at Newcastle University,
- automatic extraction of reactants, modifiers, and products involved in the yeast metabolic reactions in conjunction with researchers at the University of Manchester,
- automatic recognition and linkage of genes and enzymes associated with specific plant traits, such as biomass yield, with a team at Rothamsted Research, and
- filtering and selection of documents to provide details of kinetic parameters for reactions, in collaboration with the University of Manchester and European Media Laboratory - Research.

Education  
Evidence  
Portal

Coming Soon

visit

<http://www.nactem.ac.uk/assist>  
for more information

## Disease Extraction and Concept Association (DECA)

DECA, a one year project funded by Pfizer, concerns automatically extracting associations between concepts in the biomedical domain, such as diseases and symptoms, from collections of biomedical texts (e.g., MEDLINE™). A considerable amount of research was put into lexical disambiguation of the biomedical names. This is because storing information in the form of words can cause ambiguity, since a string of words often refers to different meanings depending on the context. Therefore, a more sensible way to organise information is by concept, where a concept has unambiguous meaning and can be associated with a unique identifier. To make text mining useful for the community of biological sciences, one crucial step is to link the hidden and ambiguous mentions of named entities in text to unique concepts in knowledge bases, namely, disambiguation.

In particular, DECA tackled one major source of ambiguity in entity mentions: model organisms. Model organisms are species studied to understand particular biological phenomena. Biological experiments are often conducted on one species, with the expectation that the discoveries will provide insight into the workings of others, including humans, which are more difficult to study directly. From viruses, prokaryotes, to plants and animals, there are dozens of organisms commonly used in biological studies, such as *E. coli*, *C. elegans*, *Drosophila*, *Homo sapiens*, and hundreds more are frequently mentioned in biological research papers. Given an article, it is often essential for readers to understand what organisms the biomedical entities (e.g., proteins) belong to, and on which organisms the experiments were carried out.

The approach to organism disambiguation in DECA was to automatically identify the species indicating words (e.g., human) and biomedical named entities (e.g., protein P53) in text, and then to judge whether the relations between them are positive species-entity ones, where a positive relation means that an entity belongs to the organism indicated by a species-indicating word. Natural language syntactic parsers and machine learning techniques were applied to classify the species-entity relations.

<b>Forthcoming Events involving NaCTeM</b>	
TBA	Text Mining Briefing Day – Systems Biology - MIB, Manchester
11-14 October 2009	17th Cochrane Colloquium - Singapore
28-29 October 2009	Workshop: Text Mining support for scholarly communications and Repositories, co-hosted with UKOLN - CBI Conference Centre, London
TBA	Workshop: Systematic Reviews, co-hosted with EPPI
16th November 2009	Text Mining Briefing Session – Medical School and British Society for Rheumatology Biologics Register - Manchester
7-9 December 2009	JISC All Hands Meeting

<p>If you would like to be involved in any of the events hosted by NaCTeM, please contact:  <a href="mailto:Sasirekha.Palaniswamy@manchester.ac.uk">Sasirekha.Palaniswamy@manchester.ac.uk</a>            NaCTeM, Manchester Interdisciplinary Biocentre, 131 Princess Street., M1 7DN, Manchester            Tel: +44 (0) 161 306 3091 Fax: +44 (0) 161 306 3099</p>
<p>This NaCTeM Newsletter was edited by Brian Rea            Email: <a href="mailto:brian.rea@manchester.ac.uk">brian.rea@manchester.ac.uk</a>            The deadline for submissions to the next issue is September 17<sup>th</sup> 2009</p>

## Recent Publications

For a list of all publications from NaCTeM visit <http://www.nactem.ac.uk/publications.php>

- Sasaki, Y., Rea, B. and Ananiadou, S. (2009). **Clinical Text Classification under the Open and Closed Topic Assumptions**. In: International Journal on Data Mining and Bioinformatics (IJDMB), 3(3)
- Piao, S., Tsuruoka, Y. and Ananiadou, S.. (In Press). **HYSEAS: A HYbrid SEntiment Analysis System**. In: Proceedings of the Fourth International Conference on Interdisciplinary Social Sciences
- Tsuruoka, Y., Tsujii, J. and Ananiadou, S.. (In Press). **Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty**. In: ACL-IJCNLP 2009
- Ananiadou, S., Weissenbacher, D., Rea, B., Pieri, E., Lin, Y., Vis, F., Procter, R. and Halfpenny, P.. (In Press). **Supporting Frame Analysis using Text Mining**. In: Proceedings of the 5th International Conference on e-Social Science
- Sierra, G. E., Cedeño, A. B. and Ananiadou, S.. (2009). **An Improved Automatic Term Recognition method for Spanish**. In: Gelbukh, A.(Ed.) Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City, Mexico, pp. 125--136, Springer
- Venturi, G., Montemagni, S., Marchi, S., Sasaki, Y., Thompson, P., McNaught, J. and Ananiadou, S.. (2009). **Bootstrapping a Verb Lexicon for Biomedical Information Extraction**. In: Gelbukh, A.(Ed.) Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), pp. 137--148, Springer
- Tsuruoka, Y., Tsujii, J. and Ananiadou, S.. (2009). **Fast Full Parsing by Linear-Chain Conditional Random Fields**. In: Proceedings of the 12th Conference of European Chapter of the Association for Computational Linguistics (EACL-09), pp. 790—798
- Kano, Y., McCrochon, L., Ananiadou, S. and Tsujii, J.. (2009). **Integrated NLP Evaluation System for Pluggable Evaluation Metrics with Extensive Interoperable Toolkit**. In: Proceedings of the NAACL HLT Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp. 22—30
- Thew, S., Sutcliffe, A., Procter, R., de Bruijn, O., McNaught, J., Venters, C.C. and Buchan, I.. (2009). **Requirements Engineering for E-science: Experiences in Epidemiology**. In: IEEE Software, 26(1), 80-87
- Ananiadou, S., Okazaki, N., Procter, R., Rea, B. and Thomas, J.. (2009). **Supporting Systematic Reviews using Text Mining**. In: Social Science Computer Review
- Ananiadou, S.. (2009). **Text Mining for Biomedicine**. In: Prince, V. and Roche, M.(Eds.) Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration, 1—11
- Sasaki, Y., Thompson, P., McNaught, J. and Ananiadou, S.. (2009). **Three BioNLP Tools Powered by the BioLexicon**. In: Proceedings of EACL 2009 Demonstration Session, pp. 61—64
- Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J.. (2009). **U-Compare: share and compare text mining tools with UIMA**. *Bioinformatics* doi:10.1093/bioinformatics/btp289