Expression of Interest: NATIONAL CENTRE FOR TEXT MINING

1. OVERVIEW
This is a consortial application involving UMIST, the Victoria University of Manchester (VUM), the University of Liverpool (UoL), and University of Salford (UoS), to host the NATIONAL CENTRE FOR TEXT MINING (TMC) (JISC Circular 7/03) as a proof-of-principle service leading to production. The service activity will be run by the national dataset service, MIMAS, based within Manchester Computing. Although a collaborative proposal involving international partners and datasets, funding is only requested for UK HEI participants.

The TMC consortium represents one of the largest concentrations in Europe of workers in bio-related information, and in bio- and general text mining. The consortium can demonstrate international reputation and contacts, industry orientation and collaboration, research track record, experience of mounting a national service through MC, and much regional support and commitment. The HEIs are close geographically which should facilitate ease of working and regular meetings, have long established working relationships, and existing joint activity. The immediate region contains a high proportion of international biotechnology and pharmaceutical companies compared to other UK areas, thus the consortium is close to much of the initial user population.

The TMC will be located in the currently under construction £28M Manchester Interdisciplinary Biocentre, largely funded by the Wellcome Trust; given the initial focus on bio-related information, this co-location is critical to successful proof-of-principle in 3 years. The eScience North West activity will install a GRID data node in this building to support the TMC. In recognition of the importance of the TMC, the UoM has agreed to establish a chair and lecturership in Information Extraction (approx. £400K investment over 3 years), whilst the UoL has agreed to sponsor a Project Manager position. The Chair appointment is expected to join the Technical Directorate.

Text mining is a set of techniques for the intelligent retrieval of textual information against complex search criteria. Text mining as a discipline derives in part from the base of data mining, natural language processing, and information extraction. The consortium has significant international presence in text mining, its base disciplines, and is the leading UK centre of activity in bioinformatics, with associated excellence in bio-computation.

The Technical Directorate has representatives from industry, from Life Sciences, Science and Engineering, Humanities, Digital Libraries, Mersey BioBank, and Manchester Computing. Close relationships between TMC and the eScience NW centre are facilitated by both ESNW directors (Goble, Brooke) being on the TMC Steering Committee, in addition, one of the TMC Technical Directorate (Keane) is on the ESNW Steering Committee. In addition, the Director of the North-East Regional eScience Centre is on the Steering Committee (Watson).

The International Advisory Panel includes major international individuals and institutions in text mining, bioinformatics, and digital libraries. Significant industrial support is outlined in the supporting letters from EBI, European Language Resource Association, IBM, Unilever, and the Amaze/Autonomy/MerseyBIO consortium.

International Advisory Panel: Prof Dr Walter Daelemans (UAntwerp); Prof Carole Friedman (Columbia U); Prof Ray Larson (UC Berkeley); Dr Clifford Lynch (Executive Director, Coalition for Networked Information, US); Prof Reagan Moore (UCSD); Prof Nigel Allinson (Science & Engineering); Prof Terri Attwood (Life Sciences); Ms. Julia Chruszcz (Director of National Data Services, MC); Prof Andy Cossins (Consortium for Post-Genome Science, UoL); Prof John Keane (Humanities); Dr Dietrich Rehboz-Schuhman (Head of Text Mining, European Bioinformatics Institute); Dr Geoff Wainwright (MerseyBIO); Dr Paul Watry (Digital Libraries, UoL).

Technical Directorate: Prof Nigel Allinson (Science & Engineering); Prof Terri Attwood (Life Sciences); Ms. Julia Chruszcz (Director of National Data Services, MC); Prof Andy Cossins (Consortium for Post-Genome Science, UoL); Prof John Keane (Humanities); Dr Dietrich Rehboz-Schuhmann (Head of Text Mining, European Bioinformatics Institute); Dr Geoff Wainwright (MerseyBIO); Dr Paul Watry (Digital Libraries, UoL).

Steering Committee: Dr Sophia Ananiadou (Text Mining, UoS); Prof Keith Bennett (Software Maintenance, UDurham); Prof Rob Beynon (Proteomics, UoS); Prof Andy Brass (Bioinformatics); Dr John Brooke (Director, ESNW); Prof Terrence Fernando (Virtual Prototyping, UoS); Prof Carole Goble (Director, ESNW); Prof Douglas Kell (Bioanalytical Science); Prof John McCarthy (Academic Coordinator, MIB); Mr John McNaught (Text Mining); Prof Norman Paton (Information Management); Dr Ray Paton (BioComputation, UoL); Prof Alan Rector (Medical Informatics); Prof Paul Watson (Director, eScience NE) Prof Jian-Bo Yang (Decision Support).

2. BACKGROUND
2.1 Service Provision
The TMC service provision will be run by the National Centre for Datasets Services (MIMAS), based in Manchester Computing (MC). The MIMAS Director (Chruszcz) is on the TMC Technical Directorate. MIMAS has worked closely with the development team at the UoL (led by Watry, also on Technical Directorate) to develop and implement national digital library services on a production basis (ZETOC, Archives Hub, IESR, etc).

2.2 Wider Context
Support in various forms has been already offered from such UK bodies as:
• North West Development Agency’s BioNoW initiative;
• North West Institute for Bio- and Health Informatics (NIBHI);
• UK Biobank: tracking 500,000 volunteers and collecting data for future research studies;
• NOWGen: North-West Genetics Knowledge Park;
• AiMes GRID Application Institute (NWDA/ERDF-funded).

2.3 Links with International Centres of Excellence
The consortium is actively engaged at the international level and closely associated with most of the leading centres worldwide. In Europe, UAntwerp provides the text mining component of the BioMinT project coordinated by VUM; UMIST has collaborated with numerous EC centres of excellence, companies and government agencies
on text mining; Salford has an on-going collaboration with the European Bioinformatics Institute and has worked closely in the past with Lion Bioscience in the EUREKA BioPATH project.

In the US, UMIST has close links with UPenn (largest recent NSF text mining grant on ‘mining the biomedical bibliome’) and NYU (large NSF grant on ‘on-demand information extraction’) through EAGLES/ISLE; the Liverpool digital libraries’ activity collaborates closely with UCal (Berkeley) (recently announced NSF grant on ‘improved search over bioscientific literature’) and UCSD, while the bio-computation activity is sponsored by Los Alamos National Laboratory; Salford has also worked with UPenn and is working with Columbia (NSF grant on ‘use of natural language processing for curation of biological databases and ontology creation’) to co-edit a special issue on text mining of the Journal of Biomedical Informatics.

In Japan, Salford has a number of links, particularly with the Human Genome Centre. UTokyo on bio-text mining, and provided the sole academic representative (Ananiadou) for a DTI OSTEMS mission investigating text mining and language processing in Japan, who was also 1 of only 4 recipients of a Japan Trust fellowship (1997).

2.4 Expertise and Competences represented in the consortium
This section presents an outline of the various expertise and competencies of staff associated with the consortium, citing only activity relevant to the bid. The existing very strong cohesion between the partners should be noted.

2.4.1 Bioinformatics and Genomics
This activity has seamlessly integrated experimentalists and bioinformaticians, with both interacting directly with information scientists. The bioinformatics activity has applied established and novel IT to the storage, access, and analysis of sequence and functional genomic data.

At the sequence level, Atwood has established and maintains PRINTS, the world’s largest manually annotated protein family database. PRINTS is used worldwide, e.g., Antwerp, Geneva, Seattle, Stanford. Automatic annotation tools for PRINTS are being developed (Atwood, Goble). Atwood leads the EC BioMinT project, which aims to provide a generic text mining tool initially to aid curation of biology databases (SWISS-PROT, PRINTS) and to generate readable prose for biology researchers, in consultation with an End Users’ Club.

Hubbard and Petrounias apply data mining for protease sequence and structure. Hubbard has other funded BBSRC projects (with Stapley and Yin) involving machine learning. Stapley has worked on predicting and extracting properties of biological entities through combining large-scale text analysis with experimental biological data and genomic information resources. His use of supervised learning over both text and biological data sources increases novelty detection, with marked success in predicting lethal toxicity of organic compounds on freshwater fish, compared to traditional methods. Recently, work has started on non-supervised learning approaches using sophisticated term and term relationship extraction (inspired by Ananiadou’s results). The overall goal is to discover strategies and methods that facilitate user comprehension of experimental data, genomic information and biomedical literature simultaneously. Stapley and Keane are investigating text mining for protein functional classification using Support Vector Machines. Paton (UoL) investigates the prediction of protein structure and function, functional genomics and proteomics.

2.4.2 Text Mining
Text mining relies on accurate and extensive ontologies. Brass, Paton and Goble (VUM) have exploited ontologies to construct TAMBIS, the first bioinformatics system to integrate disparate, heterogeneous information sources. Goble and Stevens have worked on numerous large ontology-related projects including metadata issues for the semantic web, being concerned with manual ontology construction. Rector is leading the JISC Co-ODE project on open source tools for ontology representation and management. This work is complemented by work on semi-automatic ontology construction and term clustering for ontology building (Black, McNaught, Ananiadou) automatic, funded by industry (Wordmap, Lion Bioscience). Wood and Rector are combining information extraction with knowledge representation for biodiversity informatics. Wood earlier worked on automatic compilation of taxonomic databases.

Ananiadou’s team has an international reputation in term extraction and management focusing on biomedicine, and has developed the well-regarded ATRACT system. Industrial tests by Lion BioScience on biomedical text gave 98% precision for multi-word term extraction, and 99% precision for acronym extraction. Black and McNaught have developed a mature IE engine, CAFFE, over numerous EC/industry sponsored projects. With Theodoulidis, they are currently investigating temporal text mining in the EC PARMENIDES project (coordinated by UMIST), which covers three domains including biotechnology. The addition of a temporal dimension to data impacts markedly on algorithms that would otherwise satisfy scalability and efficiency criteria, thus advanced temporal data mining solutions are highly relevant to the success of a large-scale text mining service. PARMENIDES is the only project in the world developing methods for user-oriented evaluation of text mining using the EAGLES/ISLE-ISO evaluation framework. Temporal text mining attempts to determine temporal sequences of events via temporal IE and then, via temporal data mining, to discover causal chains, or associations between events in texts and time series data. Ontology evolution over time can also be tracked by such techniques, which are thus crucial not only for finding out about synchronic time-related information, but also for tracking diachronic change in knowledge.

Work in document classification and search engine development has led to a commercial document management package QUEST under license agreement with UMIST (Allinson, Keane, Yin). This approach

1 EAGLES: European Advisory Group for Language Engineering/International Standards for Language Engineering: NB McNaught acts as one of two chief editors.
integrates document structure with statistical - text is translated into a numeric representation and statistical approaches to pattern matching are used - and semantic analysis - word meaning contained within the text, and their context, are utilised to establish information relating to the document. A virtual document contains meta-data representing semantic information including document structure, word senses/parts of speech, phrases; references to other texts, proper nouns/acronyms, concepts (semantics discovered using statistical methods). A self-organising map automatically categorises documents into topics and subtopics based on content. Clusters, representing topics, can be meaningfully labelled with the most common representative name, phrase or word in the respective document. Previous approaches using SOMs for document clustering have used a single feature vector to represent a single document in the map. A more effective approach used here is to utilise meta-data extracted from the document during the identification of sub-topics and topics. This information enables the document to be segmented based on its internal structure and constituent elements. Each segment can be used as a feature vector, enabling more detailed clustering. The sense of each word in a document segment is determined, and the frequency of usage of each word sense makes up the feature vector for that segment.

Black, McNaught, Ananiadou and Wood have long experience of natural language processing in general, covering dialogue, machine translation, parsing, generation, sublanguage, terminology, linguistic resources, IR, multilingual issues, etc. There is essential competence in multilingual issues, considering that IE, from the user's point of view, might take place from texts in any language, not just English.

2.4.3 Information Retrieval
The University of Liverpool (Watry) and UCal (Berkeley), have jointly developed a next generation online information retrieval system (Cheshire) based on national and international standards and in use by a wide variety of national service and projects throughout the UK. The software has been designed to address the need for developing and implementing advanced networking technologies required to support digital library services and online learning environments, especially those relating to the integration of digital and internet based services, digital content, and emerging applications in education. The Cheshire software and tools link a wide variety of JISC resources and extend the capabilities of the software and services characteristic of the digital library environment. JISC-funding was obtained to develop a fully threaded version of the system permitting simultaneous processing of multiple queries within the same address base. This development make the software suitable for distributed storage of databases and a distributed processing in a GRID-like architecture.

myGrid (Goble) seeks to extend the Grid framework of distributed computing, producing a virtual laboratory workbench to serve the life-sciences community. Different kinds of functional genomic data need to be integrated to generate new generalizations about biology. GIMS (Oliver, Brass, Paton, Goble) is an object-oriented database warehouse for sequence and functional genomics data. Kell and Brass are involved in providing databases of transcriptome and proteome data for a data warehouse which will be linked with genomic and annotation data from the John Innes Centre. Kell’s work also interprets complex metabolome data via evolutionary computing.

2.2.4.4 Accessibility, collaborative working, visualisation
A service must respond to end-user requirements and, in modern interdisciplinary frameworks, to collaborative working and the need to visualise complex results in meaningful ways. Salford and UMIST are highly active in these areas, via work on HCI, virtual reality, collaborative engineering, visualisation and future workspaces (Fernando), and on scenario-based requirements engineering, HCI and systems integration (Sutcliffe).

The UMIST Centre for Rehabilitation Engineering, Speech and Sensory Technologies (Blenkhorn, Evans) have some 25 systems for accessibility under license and were responsible for the Narrator screen reader commissioned by Microsoft, and for a head-operated mouse that gained an Innovation in Education Award. Sponsors include Guide Dogs for the Blind Association, and the EC (TeDUB: understanding of technical drawings for the blind; DUMAS: on adaptive speech interfaces). In addition, all MIMAS services fulfill the accessibility requirements as detailed in the SENDA legislation and apply W3C guidelines.

2.2.5 Analysis of existing commercial text mining software and process
PARMENIDES conducted (April 2003) a technology assessment and comparison of text mining tools, including commercial, and of tools representing supporting components (document collectors, document warehouses, data mining subsystems, visualisers). There is thus detailed knowledge in the TMC consortium of commercial tools falling into three broad classes of architectures: component-oriented, file-oriented and database-oriented, namely e.g. Megaputer’s TextAnalyst, IBM’s Intelligent Miner for Text or ClearForest’s suite, and Oracle Text or Thunderstone’s Textis. Current tools lack integration of different techniques, scalability, and temporal processing.

Recent reviews of bio-related text mining are given in Nature Reviews: Genetics, Drug Discovery Today, and Briefings in Bioinformatics. The conclusions are that much current activity concentrates on restricted areas such as cellular pathways and protein interaction, that complex fact extraction and inference is largely missing, that there is little work on exploitation of multiple documents (hardly any on documents in multiple languages) and non-text sources (integrated mining), that manual ontology construction, while accurate, is slow and semi-automatic construction is a key area of investigation, that there is a lack of a common evaluation framework, and that there is little attention paid to temporal aspects. Recent NSF-funded projects have begun to address these issues. All reviews point out that information extraction systems do not handle terminology well – a critical issue for all domains. Attempts range from the linguistically naive (EDGAR), to use of terminology grammars and morphosyntactic cues (PROPER, GENIES, EMPhIIE, PASTA) where only terms belonging to specific terminological classes (e.g. enzymes, proteins) are recognised. All these IE systems rely heavily on looking up static existing resources (UMLS, GeneBank) to confirm as terms identified in the text thus typically fail to
handle previously unseen terms where all elements are novel; (as usual in dynamic domains. There is a general failing to account properly for multiword terms (the vast majority) and for term variants (acronyms, reductions).

A major focus in the assessment has been the University of Sheffield’s GATE (General Architecture for Text Engineering), a popular framework for research on information extraction (IE) which offers plug and play and communication via a common database. Tests revealed that GATE did not scale well when large amounts of documents were presented for analysis. Furthermore, the single span annotation format represented in the database is over-restrictive for advanced IE which needs to refer easily to disjoint spans. The GATE JAPE formalism also compares poorly with other state of the art formalisms such as UMIST’s CAFFE. JAPE is highly procedural (Java methods are used directly in rules to construct and manipulate representations), difficult to learn without programming knowledge, lacks a direct negation operator, and does not permit a context to be reused in the same grammar. In its procedurality it is akin to TAI’s VisualText NLP++, formalism: neither are likely to be of use in a national service where users are likely to want to develop their own grammars for their own texts. This discussion does not preclude consideration of machine learning based techniques for IE stages of text mining. Any reasonable service should offer both rule-based and machine learning based solutions. In general, the favoured route here is to eschew a GATE-like framework in favour of a web services model that communicates through a common annotation standard - UMIST is involved in the development of such a standard.

With increasing size of scientific literature, incremental text mining, building on previous and intermediate results, will become significant. Hybrid text mining, combining results from different types and sources of data (e.g. experimental results and annotated text) is likely to develop as the investigative paradigm in most domains.

3. TMC ACTIVITY

The depth of existing activity means this consortium is well-placed to develop a national text mining centre, given its innovative work on information retrieval; hybrid, integrated text mining; temporal semantics; incremental text mining; complex fact extraction and inference; integration of techniques.

3.1 Technical Activity

Given the assessment in §2.2.5, it is essential that TMC addresses identified limitations of text mining: e.g. a common evaluation framework; temporal semantics; incremental text mining; scalability in terms of document number, size, and complexity; terminology; complex fact extraction and inference; integration of techniques.

3.1.1 Software Development

The software development will integrate core text mining components, to augment these with the results of recently-completed or current research activity, and, where appropriate, to increase their robustness and efficiency to provide a high-quality service. Existing consortium packages that have been identified as most mature include: Cheshire, ATRACT, CAFFE, and QUEST.

The Cheshire information retrieval (IR) system is to be enhanced to allow integration of a component to meet the IR needs of text mining, with particular work on advanced indexing and retrieval, focusing on metadata, on improved index term weighting, on search interfaces, and on ontology management. The SKDL data mining toolkit will be integrated to associate entities, such as parsed genome data and abstracts with bioscientific texts and bibliographic data. The capabilities of myGrid will be integrated into this framework. The primary advantage is that Cheshire will be able to support hybrid text mining (e.g. form a journal and from DNA) in a transparent and efficient manner.

The ATRACT term extraction, clustering and management system requires sophisticated and reliable filters are required to allow extraction from particular user-defined term mining spaces, e.g. a user may want to find only new terms referring to proteins, and on implementing techniques to allow incremental extraction.

The CAFFE Information extraction system requires further integration of ATRACT results and ontological knowledge is required. A systematic method to ‘cache’ intermediate results will enable incremental text mining. An annotation editor would support user creation of annotated textual resources and user validation of information extraction result. A rule editor would support users in elaborating their own rulesets. Both editors would support access to ontologies.

The QUEST text mining semantic analysis component will be extended to address proximity metrics and to improve word sense disambiguation results with latent semantic indexing, by using more than one document.

The intention is to develop a component-based architecture to support users in defining text mining scenarios through appropriate composition of components running over user-supplied or remotely-accessed resources. This should allow 3rd party and user-specific components to be integrated. This standards-based model will ensure the widest possible outreach and utility of the service. At all times, development will consider accessibility aspects.

3.1.2 Evaluation Framework Development and Evolution

The needs of users of text mining are varied and vary over time even for one user. There are many text mining tools and components: generic/specific, customisable/inflexible. The task of specifying a text mining scenario is complicated by the many tools on offer. There is currently no standard means to evaluate text mining tools. Users of a text mining service, which may offer a range of tools, would significantly benefit from a standard framework to allow them to specify the characteristics of their task such that an evaluation of available tools is possible.
Indeed, it is desirable for users to access, as a service, an evaluation framework that allows them to compare functionalities and characteristics of text mining tools whether or not the service actually provided these tools.

Such a framework is being developed within the UMIST-coordinated EC PARMENIDES project, based on work within the EC/NSF ISLE project under the long term EAGLES initiative which extends ISO/IEC 9126-1 and which is carried out in cooperation with ISO TC37/SC4. A similar type of evaluation framework was developed in ISLE for machine translation tools (there being many types of machine translation and many differing user needs of MT) which has proved highly popular (see http://www.isasco.unige.ch/projects/isle/femti/). It is proposed to offer and further develop, as part of the proposed TMC service, an evaluation framework for text mining that will conform to the PARMENIDES evaluation model or appropriate eventual ISO elaboration thereof.

3.3 Service Activity
3.3.1 Service Provision Outline Timetable
Given the existing software in the consortium, in year 1 the most innovative and/or stable of these systems will be made available, with technical advice rather than service-level provision. Consideration will be given to commercial software offerings from such as IBM, Autonomy, Amaze, and IBM. Between 18 months and 2 years, a prototype service-level provision of the evaluation framework and the developed software will go on-line. In year 3 a more comprehensive text mining service will evolve thus evidencing proof-of-principle.

3.3.2 Outreach into new areas
There is much competence within the consortium in medical- and chem-informatics hence these will offer initial spread of impact. In engineering, there is identified potential in areas such as fault analysis – maintenance reports within electrical power systems – and product design engineering – tenders, specification, description.

A major challenge is to pull through the service provision into the Humanities sector. In support of this there is Humanities presence on both the Technical Directorate and the Steering Committee. Feasibility studies in year 1 of the TMC will investigate open government, law, decision support, design science, software maintenance, and web mining. The most promising of these will be developed with larger-scale exemplars planned for years 2 and 3. Close collaboration with the UK eSocial-science centre will be critical for user acceptance and dissemination. It is further crucial to communicate progress on all fronts to industry, and to be informed by their current activities.

4. TMC ORGANISATION
4.1 Staffing in the Centre
As outlined in the call, the expectation is for 6 staff. One will act as TMC Manager, with a further post providing admin support for staff and the various committees. Other staff will focus on service development, deployment, and provision, user support and training, technical advice, software development, and research pull-through and feedback to domain workers – given the initial focus, one or more staff will be in bio-related areas. As data access and management is critical, it will be a designated responsibility. The Manager will act as the crucial link to the business development facilities within the HEIs; business development is thus regarded as a strategic core activity.

4.2 Operation of the Centre
TMC day-to-day activity is the responsibility of the Operational Committee, including the Manager, and nominated representatives of the Technical Directorate, the Steering Committee, and the User Forum. The Technical Directorate will liaise regularly to consider progress and influence direction. The Steering Committee will meet twice yearly – following consultation with the International Advisory Board – to influence overall direction. The International Advisory Board will input to the Technical Directorate at any time, and formally twice yearly.

4.3 Dataset Access and Facilitation
The University of Manchester, through MIMAS, CSAR, and BioBank, host significant data collections. The TMC will seek appropriate access to these resources and other collections to develop its work. For example, such collections might include zetoc the British Library Electronic Table of Contents hosted at MIMAS and BioBank.

Users will be encouraged to annotate a subset of their datasets to identify relevant entities/facts, and their utility. This will provide reference corpora for evaluation and standardisation, exemplars to build best practice within a domain, and useful resource for technical progress in machine learning and rule-based approaches.

4.4 Dissemination and Outreach
The dissemination and outreach activity will borrow from the established MC service processes, such as virtual user groups, newsletters, training, web forums etc. In the first 18 months, TMC will inaugurate an annual Text Mining Workshop with complementary themes: service provision; user requirements and experiences; technical developments. This should enable interested parties to engage very transparently with the service provision plans. Opportunities to host international conferences will be pursued. Regular seminars detailing user experiences and technical advances, and user technical guidance via helpdesks will be organised as the service matures (years 2, 3).

4.5 Associated Technical Staff
Text Mining: Mr Bill Black; Prof Kit Dodson; Dr Goran Nenadic; Dr Iliai Petrounias; Dr Babis Theodoulidis; Mr Christos Tjortjis; Dr Hujun Yin; Bioinformatics: Dr Simon Hubbard; Dr Ben Stapley; Dr Robert Stevens
Natural Language Processing: Prof Allan Ramsay; Prof Harry Somers; Dr Mary Wood
Assistive & Interactive Systems Design: Prof Paul Blenkhorn; Dr Gareth Evans; Prof Alistair Sutcliffe