

Clinical E-Science Framework (CLEF): Case for Support

1. The nature of the research challenge and aims of the proposal

CLEF focuses on generic methods for the capture, processing, and dissemination of information about cancer patients and their care using electronic health care records integrated with support for clinical and basic research in the biosciences. It will incorporate this information in formal knowledge management systems within the e-Science/Grid framework. It will result in tools to support faster, more effective organisation of clinical trials, improved identification of eligible patients, and rapid screening of hypotheses. It will provide the foundations for cooperative multidisciplinary care, evidence-based practice, and quality assurance.

There is growing recognition that challenges in health informatics research are central both to the modernisation of health services and to a successful strategy for clinical research to exploit our new knowledge of genetic and genomic processes. Clinical e-Science faces special barriers in capturing information and assuring its quality — problems also central to realising the potential of information for improved healthcare.

CLEF aims to embed the full information cycle — capture, integration, interpretation, and use — within practical clinical systems while observing strict ethical and legal requirements for confidentiality and consent. To achieve this ambitious but crucial goal requires addressing organisational issues in information governance and solving technical problems in language technology, information integration and information presentation. In particular CLEF seeks to unlock the information now held only as text in reports and summaries and therefore largely unavailable for analysis. CLEF also seeks to address the scale and complexity of clinical information and to link it to emerging genetic and genomic information. To do so requires addressing the problem of metadata and knowledge management using very large ontologies — hundreds of thousands of densely interconnected concepts. CLEF will make integrated clinical and genomic information available for clinical research through novel tools for collaborative work and clinical e-Science. The same technologies will provide the core for novel patient care systems to serve integrated multiprofessional teams and patient communities.

The specific objectives are to produce:

- Agreed policies on information governance — confidentiality, access, authentication, and consent — implemented by security measures within the technical framework of the Grid and operational NHS systems
- Tools for information capture, particularly language tools for capture from text and dictated reports, integrated in a broad strategy which improves the quality of information held
- An integrated repository of clinical information serving the SW and NC London Cancer Networks, including genomic information where available, serving clinical research and drawing its information from systems which directly support patient care, and served by tools suitable to the scale and complexity of clinical knowledge
- Tools to access and present the information in the repository intuitively, often in generated language, and to integrate it with other information in the e-Science network to support linked clinical and bioscience research and to make the cycle of clinical experiments and trials faster, more efficient, and more effective
- Evaluation and demonstration of the tools and repositories in practical scientific investigations

CLEF takes advantage of a unique opportunity to build on three major projects: a) Nuffield-funded work of the Judge Institute, Cambridge on ethico-legal requirements for research use of clinical information; b) Integration of healthcare records in the Royal Marsden NHS Trust and the North Central and South West London Cancer Networks; and c) the EPSRC funded *MyGrid* project on *Directly Supporting the E-Scientist*(www.mygrid.org.uk).

CLEF is a three year project but conceived as the core of a long term programme of research. The collaboration has two centres — a) A clinical centre around University College London Centre for Health Informatics and Multiprofessional Education (UCL/CHIME) linked to the Royal Marsden NHS Trust, the North Central and South West London Cancer Networks, and the Judge Institute for Management studies at Cambridge, and b) a technological centre around the departments of computer science at the universities of Manchester and Sheffield and the Information Technology Research Institute (ITRI) at University of Brighton.

CLEF will therefore be embedded within the everyday large scale world of service delivery — a necessary prerequisite to demonstrate that it can be widely disseminated in the NHS. Clinically, it will draw on leading cancer research units spanning the whole cycle of information management from genetics and structural biology to multicentre clinical trials and multiprofessional information systems for patients in primary care. Technologically, it will draw on leading centres in e-Science/Grid and language technologies contribute to health care and clinical research priorities being incorporated in the emerging e-Science/Grid framework.

CLEF will establish the nucleus of a public domain Clinical e-Science/Grid Community for the UK, closely linked with existing international partnerships and public domain initiatives in the field of health informatics. It will draw

on, and contribute to, existing open knowledge and software resources such as *OpenGALEN* (www.opengalen.org), *OpenEHR* (www.openehr.org) and the Gene Ontology (www.go.org) and will be open to partnerships across the whole of the e-science programme. Although focused on cancer, CLEF will be informed by the many research programmes of its participants in clinical fields such as in cardiovascular disease, genetics and neuroscience.

Scenarios: The long term goals of the CLEF programme are illustrated by three scenarios. Not everything can be achieved in a single project, but CLEF aims to remove key barriers to their realisation.

1.1. Faster, more cost effective clinical trials and experiments linking clinical and genomic information

The key long term goal is a step-change in the speed and effectiveness of clinical research and drug discovery whilst respecting patients' rights to privacy and informed consent.

A clinical researcher is attempting to confirm suggestions that part of the variation in response to a treatment protocol may be related to genetic markers and patterns of genomic expression and that the risk of certain adverse events can be predicted from comparison of processed pre- and post- treatment radiographs.

The researcher is registered with the Repository Custodian Authority to have automatic access to fully anonymised pooled information, but not full text individual records. Although the critical clinical details from discharge summaries and radiography reports were recorded only in dictated text, the mark up by the information extraction software is sufficient for preliminary analysis. The presentation is in generated outline-style natural language which the researcher finds natural to read and use. Filtered links to Medline, Cancer Net, policy material from NICE and Cochrane reviews are linked automatically, along with links to a unified view of the genomic literature provided by MyGrid. The e-Science Workbench dynamically builds a browsable dossier of current mainstream and grey literature linked to the electronic healthcare records of the patients under review.

Not all data is strictly comparable, but the ontology allows decisions to be made concerning which items can be safely aggregated for preliminary analysis — the clinical equivalent of an “in-silico” experiment. The researcher applies to the Repository Custodian Authority for more detailed access including full text of reports and records for selected patients, all but one of which have given prior consent for their data to be used in this way. In parallel the researcher contacts networks of clinical and molecular biologist colleagues involved in relevant virtual e-Science communities.

A study is quickly formulated, largely from existing protocols managed by the Clinical e-Science Workbench, and is sent to the ethical committees for fast track approval. Eligible patients are quickly identified and recruited via the repository and its links to the electronic healthcare record system.

The geneticists, meanwhile, note that a related effect is known in an animal model, and link to the MyGrid resources to follow up the detailed molecular biology across the range of genomic and proteomic resources.

The early results of the clinical study indicate that these radiographic and genetic markers are relevant, and the information is added to routine clinical guidelines which are widely disseminated through the links to NICE, the Cochrane Collaboration, and the Cancer Networks. It is also arranged that the relevant image processing and information extraction from the radiology reports be available in real time for clinic use — an application which requires more computing resources than NHS hospitals would normally have at their disposal — which is therefore arranged to be performed via the Grid.

1.2. More effective tailoring of care to individual patients

The same basic scenario as above is equally relevant to clinical governance and quality assurance. In addition, where it is unclear whether the evidence behind evidence based practice generalises to the patient at hand, there is major potential clinical value in complementing standard guidelines by the experience of similar patients [12], *e.g.*:

A clinician is attempting to apply an evidence based protocol to a complex patient who would have been excluded from the studies underpinning the protocol. The clinician is concerned about possible unexpected outcomes because of several intercurrent diseases. The clinician queries the repository using the relevant patient details extracted and anonymised automatically from the patient's electronic healthcare record and searches for similar patients in the repository using a combination of navigation and examples. A group of six similar patients is identified who have been treated in analogous ways. On the basis of their clinical histories the clinician decides to modify the protocol before treating the current patient. The clinician also obtains the patient's consent to add their history and subsequent clinical course to the repository.

1.3. Improved quality clinical information and reduced workload for clinicians

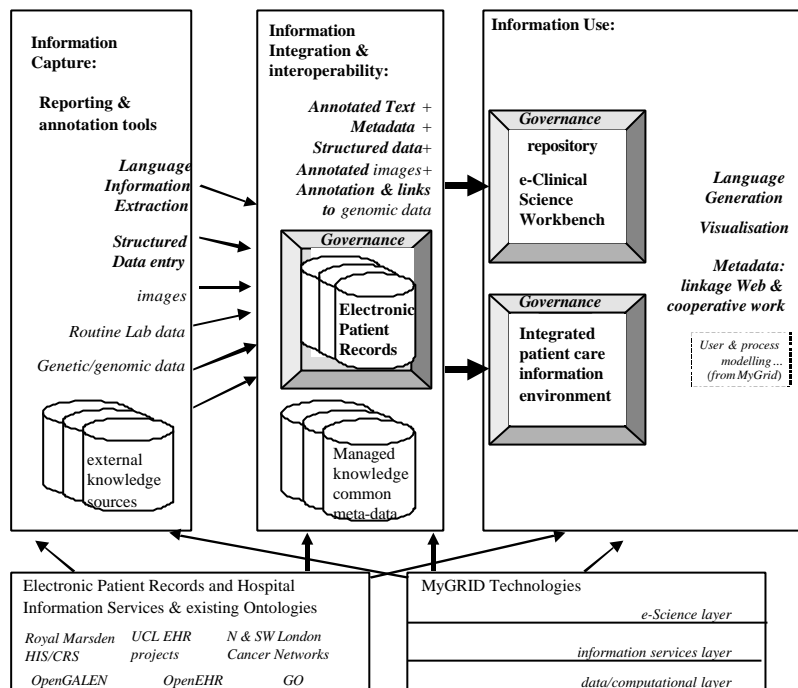
Collecting clinical information for both trials and care should be moving towards controlled use of structured text:

Having seen the patient, the clinician dictates the results according to a simple template shown on a combined PDA and dictation device. Alternatively the template might be on a simple card. The dictated information is

transcribed and key information extracted automatically. If the extraction software indicates high confidence, validation is delayed until the next time the patient is seen. If not, text is referred back to a data manager and by them, rarely to the clinician. The detail and complexity of information extracted has increased steadily since introducing the system and now covers most clinical details

Outline of architecture: The relation of the main technologies to be developed in CLEF with the underlying Electronic Healthcare Record and MyGrid technologies is shown in Figure 1. The specific contributions of CLEF are shown in bold to distinguish them from information and data imported from collaborating projects.

Fig 1: CLEF technologies and relation to Electronic Patient Records & MyGRID



2. The contribution the proposal makes to national needs

2.1 E-Science and BioScience Priorities

Linking genotypic and genomic information to phenotypic and clinical information is widely seen as the next key step in realising the potential of genomic research and a key task for e-Science. Good quality clinical data integrated into the e-Science framework is a prerequisite for achieving this aim.

2.2 NHS priorities

The National Health Service Plan and its *Information for Health* Strategy [11] looks forward to a time when electronic health care records, knowledge management services and decision support systems, built on a common government information infrastructure, will transform clinical information management and patient care within the NHS. However, early findings from implementation of the NHS Plan through the Modernisation Boards have highlighted the mixed picture of data quality and information management. Information capture lies at the heart of this problem and has been recognised as the central problem of health informatics for a decade. This problem goes to the heart of both healthcare provision and biomedical science.

Cancer is the topic of one of the three top NHS priorities and the subject of a major National Service Framework Agreement under the NHS Strategic plan. The project links closely to major efforts in the London Cancer Networks and will be a significant contribution to making it easier to coordinate the efforts of the many groups involved in cancer and related genomic research.

2.3 E-Science Priorities: Knowledge Management and Common Mark-up

Common standards for meta data for the “semantic web” are key concerns for e-Science[3]. While medical research has traditionally led the way in analogous areas — *e.g.* the MeSH headings for bibliographic retrieval and the International Classification of Diseases for epidemiology — it is in danger of falling behind in the Web and Grid generation of metadata. CLEF will foster common meta data standards for the semantic web and Grid in medical and bioscience research. Participants are active in the relevant committees of CEN, ISO, HL7, and MyGrid is active in Web standards organisation, W3C. An early deliverable will be a series of workshops on common metadata and indexing standards to include key players, *e.g.* the British National Formulary, BMJ

Evidence, the NHS Information Authority, the Cochrane Collaboration, and the National Electronic Library for Health.

2.4 Making it possible for clinical applications to use the Grid

CLEF will bring to MyGrid its focus on governance, language technology, links to operational NHS systems, and the scale and content of clinical ontologies and meta-data required for clinical applications — see 3.1 below.

3. Relationship to on-going national/international e-science activities

3.1 CLEF, MyGrid, and e-Science

MyGrid — Directly Supporting the e-Scientist — is developing the fundamental Grid based technologies for supporting personalised collaborative e-Science. MyGrid focuses on basic biological science. CLEF extends MyGrid to embrace clinical research and links to electronic healthcare records. CLEF takes from MyGrid its emphasis on process, collaborative working, technical interfaces to Grid technologies, and low level security framework. CLEF brings to MyGrid the links to electronic healthcare records, links to national programmes on information governance and the use of clinical information for research, a broadened meta data framework encompassing clinical as well as biological information, and tools to cope with very large ontologies. CLEF will share and reinforce MyGrid's work on language technology and information extraction.

A key benefit of CLEF to the clinical research community will be to ensure that the requirements of clinical research and clinical practice are recognised and taken into account in the development of the UK and international Grid infrastructure.

3.2 Links to other e-Science Projects

The Judge Institute is a key partner in a major international project on Privacy and Secondary Use of Data in Health Research funded by the Nuffield Foundation. Manchester is the hub of the Grid and BioScience activities in the Northwest. In addition to MyGrid, Manchester is also home to a major BBSRC funded programme — *Platform for Annotation, Robust Analysis, Data Integration and Genome Management — PARADIGM* (Prof Attwood) and is bidding to become the *National Genetics Reference Laboratory* (Dr. Elles). The *Manchester Information and Associated Services (MIMAS)* is the major Grid centre for the Northwest and the home of *E-Science North West* whose special focus is biomedicine (see also Section 16). Manchester (Prof Taylor) and UCL are both participants in the IRC: *From Medical Images and Signals to Clinical Information (MIAS)* which has e-Science funding for a complementary project, also using the MyGrid architecture, to build a distributed repository of medical images and associated clinical information. A collaborative project with the *Dependability IRC* will extend this work by exploring technical and organisational issues in sharing clinical information on the Grid. Other E-Science and Grid projects led by or participated in by the University of Manchester include *EuroGrid*, *RealityGrid* (experiment visualisation and steering); *Geodise* (fluid dynamics); *DataGrid* and *GridPP* (PPARC)

UCL is a member of the London e-Science Consortium which brings together activities at major research-based Universities within the London region and the regional e-Science centre based at Imperial College. The UCL Grid Forum, brings together research teams active within PPARC, BBSRC, NERC, EPSRC and MRC e-science domains, as well as within the EU funded Grid programme (eg *EGSO*, *GRIDSTART*, *DATATAG*, *DATAGrid*). It also participates in EU pilot development and health care middleware demonstrations (*6NET* and *6WINIT*).

CLEF will complement and will collaborate with the concurrently proposed project led by Professor Richard Begent at UCL, *An Information System For Development of Cancer Therapy*, in three ways: a) common ontology development, b) analysis of clinical data and c) data integration and data mining.

Professor David Ingram is on the advisory board for IT to the joint MRC-Wellcome *BioBank* large cohort study.

Manchester and CHIME have close links with the European Bioinformatics Institute (EBI) which also participates in *MyGrid*.

4. The Structure and management of the research team

The project will be organised around two centres with technical direction at U Manchester (Prof Rector) and clinical direction at UCL (Prof Ingram), under the overall direction of Professor Rector. The project brings together the expertise in information integration, e-Science and the Grid from U of Manchester with the 'real world' experience of large scale Health Service systems at UCL and the Royal Marsden NHS Trust. The collaboration grows out of links established through the UK Institute of Health Informatics and the NHS Academic Core Group.

A formal project board will consist of the heads of the six participating centres. Day to day project management will be in the hands of a technical manager based in Manchester. A steering group will be constituted of the industrial and NHS collaborators to ensure dissemination. Major events will include four major workshops plus bi- and tri-lateral visits. All sites have video conferencing facilities which will be used for check point meetings.

Manchester has played a critical role in establishing the role of meta data and ontologies in knowledge management both in medicine (Prof Rector) and in Biosciences and beyond (Prof Goble) as well as in developing the current de

facto software tools and environments and participating in the W3C in OIL [4] (www.ontoknowledge.org/oil) and DAML+OIL [8]. Manchester has major BioInformatics programmes and is the focus of the MyGrid work in BioInformatics including TAMBIS, PRINTS, reformulation of the Gene Ontology in DAML+OIL. It is bidding concurrently to become the National Genetics References Laboratory and for projects linking genotypic and phenotypic information. Involvement in the MIAS IRC (Prof Taylor) links CLEF into a major MRC/EPSRC initiative, creating the potential to extend the work to the important domain of medical images and signals.

The collaboration between the computer science departments at Manchester and Sheffield is well established within MyGrid and other projects. Sheffield has an international reputation in information extraction and is playing a role in MyGrid complementary to its role in CLEF, thereby adding to a critical mass. ITRI/U of Brighton brings to the collaboration extensive national and international collaborations on language generation.

The collaborations between UCL/CHIME, the Royal Marsden Hospital Trust, and the North Central and South West London Cancer Networks are well established and ongoing. UCL/CHIME has the strongest base in the UK in Electronic Healthcare Records in the network of projects related to OpenEHR (www.openehr.org). The Royal Marsden (Dr. Milan) has the most advanced cancer hospital information system in the UK. UCL Clinical Research Network spans cancer research in seven NHS Trusts (Professor Schapira), and the NoCTeN Primary Care Research network of UCL (Professor Wallace). It will also benefit from the close presence within UCL of the MRC (Professor Derbyshire) and CRC (Dr Ledermann) national clinical trials centres. Professor Begent's research group will provide CLEF with extensive clinical research datasets, spanning all data modalities, from genomic data to images and narrative clinical texts. CHIME has projects with the Wellcome Trust and the Collaborating Centre for Community Control of Inherited Disorders — WHO Geneva — within the Human Genetics Programme and the work of the WHO Haemoglobin Disorders Expert Working Group and on the International Human Genome Variation Society (Formerly HUGO Mutation Database Initiative).

Both Professors Rector (Manchester) and Ingram (UCL/CHIME) have experience in running large EU sponsored collaborations.

5. Detailed plans for proposed partnerships with Industry

A steering committee of industrial sponsors has been organised comprising representatives from the medical and general IT communities, the pharmaceutical industry, publishing and the NHS Information Authority. Each will provide consultancy, and Sun Microsystems and Oracle are providing deep discounts on their products. (See attached list of collaborators and letters of support.) CLEF will complement and build on existing partnerships of UCL with Oracle, ObjectStore and iSoft in the South West Devon ERDIP project. MyGrid has extremely strong industrial links which overlap with CLEF's and provide additional routes to exploitation and dissemination. Oracle and Sun Microsystems are providing substantial discounts and free consultancy to CLEF, and Sun is a major supporter of MyGrid.

6. Plan of research with clear targets and milestones

CLEF is organised conceptually in two layers: a) The clinical exemplars which provide the global demonstrators, requirements and evaluation criteria; and b) the individual technologies each of which have specific requirements, objectives, evaluation criteria of their own. Each technology workpackage is designed to have its own progressive migration pathway and milestones. There will be a global architecture, developed in its own workpackage, to link the Electronic Healthcare Record (OpenEHR) to MyGrid, but each technology provider will provide for integration with that global architecture. Similarly each technology provider will undertake its own technical evaluation in addition to the global evaluation. The project has four main phases as shown below which are distributed amongst the workpackages as detailed in workpackage descriptions and the attached Gantt chart.

Time	Phase	Milestones
m1-m6	Initiation and consensus	Governance policies; Requirements and evaluation criteria. Architecture consensus and assessment of MyGrid components; identification of corpora
m7-m18	Prototypes & Preliminary integration	Integration of EHCRs in London, Integrated ontologies and metadata; preliminary population of repository; first user interfaces and integration with MyGrid for the Clinical e-Science Work Bench; information extraction of diagnosis and grade for selected cancers; text generation for knowledge engineering; Security features.
m19-m30	Build of integrated systems	Revised version of EHCRs; Integration of EHCRs and repository; Integration of ontologies and repositories with other cancer network sources; extraction of treatment intentions and other key information; generation of end user reports
m31-m36	Installation & evaluation	Installation in all clinical sites; demonstration of use for patient recruitment for clinical trials; assessment of quality of information extraction and reliability measures; evaluation of users' responses to e-Science workbench and generated language.

WP0 Project Management and Integration with MyGrid (Lead: Manchester)

The project management will provide for overall day-to-day coordination and will also be responsible for managing the consensus on a common architecture and harmonisation with MyGrid.

- T1. Activation of advisory and steering committees, workshop schedules, and collaborative infrastructure
- T2. Coordination with MyGrid and consensus on overall architecture and integration standards
- T3. Consensus on global requirements and evaluation criteria across workpackages
- T4. Revision and management of workplan, milestones, and deliverables
- T5. Management of global iterative and summative evaluations

WP1 Clinical Exemplar and global evaluation

WP 1.1 Integrated Cancer Care and Research e-Science & e-Healthcare environment for North Central and South West London and the Royal Marsden NHS Trust (Lead: UCL)

The cancer networks in NC and SW London provide an opportunity to test the proposed developments in a broad operational context. The Royal Marsden NHS Trust (RMT) has an advanced Hospital Information System/Electronic Healthcare Record (HIS/EHCR) with an established Clinical Research System (CRS). In parallel, initiated by the national ERDIP programme in cooperation with UCL, an Electronic Healthcare Record is being implemented in the North Central London Cancer Network. The software from these two projects is complementary, and will be linked and made available to both NC and SW London Cancer Networks to provide an integrated information system supporting patient care, quality audit, and repositories for e-Science linked to key research centres in London and elsewhere.

CLEF will provide an integrating force for these developments and tackle the key problems of information capture and governance and make the results part of the national e-Science resource. Organisational issues are a major barrier to extending these programmes and will be the responsibility of the UCL team. The project will progress a) by progressive integration, and b) by progressively increasing the amount of information extracted from text and working with the data managers and clinicians to establish the optimum balance of manual and automatic extraction of information and the optimum level of structure for dictated texts.

Major milestones will be:

- T1. Implementation of a common database for SW and N London Cancer Networks with manual data entry by information managers plus links to the EHCR.
- T2. Implementation of network-wide common electronic clinic and discharge summary system using dictation/transcription linked to the cancer database aimed at a) patient care b) clinical audit c) pathways for cancer care
- T3. Implementation of controlled anonymous link to research repository
- T4. Deployment of e-Science e-Health workbench (See WP 2.6) aimed at a) identification and recruitment of patients for clinical trials, b) evaluation of hypotheses; linkage to genetic and genomic resources via MyGrid.

WP 1.2 Global Requirements Analysis and Evaluation (Lead: UCL)

One of CLEF's guiding principles is that requirements analysis and evaluation form a single task which runs throughout the project, with evaluation criteria continually reviewed in the light of experience. Requirements and evaluations are required for:

- T1. Use of repository and Clinical e-Science workbench for identification and recruitment of patients for clinical trials within the Royal Marsden NHS Trust
- T2. Use of repository and Clinical e-Science workbench for evaluation of hypotheses within the Royal Marsden NHS Trust
- T3. Use of repository and Clinical e-Science workbench for quality assurance within the NC and SW London Cancer Network
- T4. Management and enforcement of information governance

WP2 Methodologies and Technologies

WP 2.1 Information Governance: Models of access, authentication, confidentiality and security (Lead: Judge Institute of Management Studies, Cambridge)

This work will be undertaken in conjunction with the Nuffield funded project on "Privacy and Secondary Use of Data in Health Research". It will include the establishment of the policies for access, authentication, and use of the clinical e-Science repositories and organisation for the Repository Custodian Authorities which will oversee that use. In cooperation with the Nuffield project, a workshop of critical stakeholders will be held early in the course of the project.

- T1. Requirements for handling of patient health record data within CLEF, and for the record repository, drawn from the experiences and frameworks of the Judge Institute of Management Studies, Nuffield Trust, OpenEHR, the Lowrance Report, Caldicott and Data Protection legislation

- T2. Security architecture for CLEF, established and specific Grid technologies
- T3. Access control policies for the CLEF record repository
- T4. Evaluation of data flows within the CLEF project with respect to confidentiality and anonymisation

WP 2.2 Information Extraction (Lead: Sheffield)

Well-founded clinical studies require access to extensive, fine-grained data about individual patients. The bulk of this information is held in textual form in clinical reports, e.g. discharge summaries, radiology and pathology reports. Manual analysis of sufficient volumes of this data is clearly impractical; advanced language processing technologies now support the extraction and/or automatic markup of key types of information in electronic documents. Experience elsewhere, as well as by the collaborators, indicates the suitability of clinical texts for information extraction, e.g. [10].

Using generic information extraction (IE) technology developed at Sheffield with on-going EPSRC support since 1994 [5, 6] and specialised for work in bioinformatics applications [9], CLEF will develop tools to automatically identify, extract and markup key information in clinical reports. Specifically we shall extract the diagnosis, stage, and treatment intent from the patient summaries. This sequence provides a natural progressive development path, and the information will be invaluable in recruiting subjects for clinical trials and in clinical governance/auditing. This work is complementary to information extraction in MyGrid which concentrates on literature abstracts.

To enhance robust, high quality extraction, CLEF will explore the automatic acquisition of ontologies from medical text, using techniques pioneered by researchers at the University of Freiburg [7] who have agreed to collaborate.

- T1. Analysis of extraction requirements and text types (Sheffield, Royal Marsden)
- T2. Construction of a corpus of representative texts (Sheffield, Royal Marsden)
- T3. Adaptation of existing IE technology, including: acquisition/integration of terminological resources and domain-specific ontologies; modifications to support the text format and grammatical characteristics of clinical reports (Sheffield, Manchester)
- T4. Investigation of novel techniques for automatically acquiring ontologies from corpora (Sheffield, Manchester, Freiburg)
- T5. Experiments on linking clinical reports to the medical literature (e.g. MEDLINE) using techniques developed in MyGrid (Sheffield, Royal Marsden)
- T6. Integration of IE tools into overall CLEF workbench and support for deployment
- T7. User-based and technology-based evaluation (Sheffield, Royal Marsden)

WP 2.3 Conceptual Integration of metadata and terminologies (Lead: Manchester)

Common metadata standards based on ontologies are central both to the e-Science programme [3] and to Electronic Healthcare Records [14]. Standards for meta-data and ontologies are developing rapidly with the emergence of DAML+OIL as a key part of the W3C strategy for Semantic Meta Data in which Manchester has played a key role (www.ontoknowledge.org/oil)[4, 8]. Common metadata enables CLEF to link *Information Capture*, the many source documents for the *Electronic Healthcare Record*, and the *Clinical e-Science Repository*.

- T1. Liaison and coordination of requirements with key knowledge providers and industry. Early workshop and continuing liaison with key players and stakeholders in UK including NELH, BNF, BMJ, Cochrane, NICE, PRODIGY, NHSIA. Links to EHCR and pharmaceutical industry standards bodies (HL7, CDISC, DIACOM).
- T2. Adaptation of critical *OpenGALEN* tools, methods and ontologies to DAML+OIL and the MyGrid Framework and of MyGrid tools for very large ontologies.
- T3. Metadata for the domain: integration of ontologies for cancer including ICD-O and local clinical nomenclatures plus SNOMED-CT and/or the Digital Anatomist; harmonisation with the other E-Science efforts, e.g. the concurrent proposal for *An Information System For Development Of Cancer Therapy*.
- T4. Integration and harmonisation of ontologies and schemas with external bioinformatics and genomic resources (e.g. TREMBL, TAMBIS, GO) required by the repository.
- T5. Evaluation and quality assurance of integrated meta-data by query generation and faithful communication between systems.

WP 2.4 Clinical and conceptual integration of EHCR and construction of Clinical Repository (Lead UCL)

A major input to this work will come from developments already in hand at the Royal Marsden and UCL, based around the framework in OpenEHR

- T1. Liaison and requirements from main information providers, Royal Marsden and NC/SW Cancer Networks
- T2. Interworking of Royal Marsden NHS Trust, CHIME, NC and SW Cancer Network Information systems within the OpenEHR Framework
- T3. Metadata for EHCR interaction: integration of the Ontology driven approach used in MyGrid and *OpenGALEN* with the Object Dictionary approach used in OpenEHR and related EHCR architectures
- T4. Interworking with MyGrid Architecture and metadata
- T5. Setting up and evaluation of operation of repository and interfaces to MyGrid architecture

WP 2.5 Language generation (Lead ITRI)

Language generation will serve two purposes. Firstly, it will be used to produce reports from the integrated information and metadata held in the repository and electronic healthcare record (EHCR) for medical specialists. Secondly, it will be used to provide a natural-language window onto the knowledge base for the knowledge engineers building and maintaining it and for the clinical experts who must quality assure it. Experience has shown that raw meta data representations are difficult to understand or use [15]. The WYSIWYM (What You See Is What You Meant) technology developed at the ITRI [13, 16] makes it possible both to inspect and edit the meta-data representations through a natural language interface which users find intuitive.

T1. Requirements analysis for report generation and knowledge base access and editing

T2. Collection of lexical resources specific to the domain

T3. Implementation of generic language generation modules and WYSIWYM interface

T4. Integration with overall architecture

T5. Evaluation of reports and WYSIWYM feedback, iteratively during the project and in a final evaluation.

WP 2.6 The Clinical e-Science Workbench and applications programming interfaces (APIs) (Lead UCL)

The CLEF framework will be accessible at two levels: a) through a fully functional user interface for end users, and b) through APIs at the “information layer” of the MyGrid architecture for developers wishing to incorporate CLEF technology into their own applications. The design of the two levels brings together both user and developer requirements for integrating e-Science into the clinical environment. The work will be based on the existing developments at Royal Marsden NHS Trust and UCL linked to the MyGrid architecture

T1. Identification of major user and developer tasks and requirements

T2. Specification of API middleware layer interaction (in conjunction with WP0)

T3. User interface design

T4. Implementation of common user interface to integrated modules

7. Overall resource requirements and breakdown of costs; salaries, capital

CLEF is a multicentre partnership bringing together a diverse range of skills in a major engineering effort requiring several teams and significant management effort. Maximum advantage is taken of the synergy with MyGrid. In addition, the vast majority of the effort on organisational issues and most clinical resource is being provided independently by the Judge Institute at Cambridge and the Royal Marsden NHS Trust respectively.

Software tools and integration with MyGrid will be centred at Manchester and requires one senior computer science fellow who will act half time as technical manager plus 2 person years of computer science researcher to adapt and maintain the tools for ontology development, and integrate them with MyGrid, including migration to DAML+OIL. One full time clinical knowledge engineer is required to develop and maintain the integrating ontologies and metadata schemas which are central to the project and to provide liaison between technical and clinical centres.

The language technology effort requires two staff at University of Sheffield to form an effective team plus one staff at University of Brighton for language generation. Support for one extended visit to collaborators at University of Freiburg to investigate their techniques for ontology extraction is also requested.

Most clinical effort is being provided by the Royal Marsden, but one additional half time clinical registrar is required plus support for clerical effort to scrutinise the anonymisation of texts.

Two health informaticians at UCL are required to develop the user interfaces for information capture, the clinical repository, and the e-Science workbench and the integration with the existing medical record and hospital information systems.

One-half time management research fellow is required under subcontract to the Judge Institute for work specifically related to CLEF and to coordinate interaction with ethical committees.

Support for travel and workshops plus major conferences in each discipline is requested.

All networking and major computing resources will be provided by the host institutions. However, because of requirements for security and confidentiality, one secure server for each of the major sites is required. One basic PC type workstation is required for each research staff.

8. Plans for making databases interoperable

CLEF is aimed at developing interoperable, shared resources within the Grid framework. Communication with EHCR and other NHS services will use existing standards which are rapidly converging on HL7 and the OpenEHR architecture, in which Manchester and UCL respectively play active roles.

A major goal of the project is to promote and participate in the development of common ontologies which span clinical, biological and genomic information and to provide common interfaces to multiple resources as has been demonstrated by the TAMBIS [1] (img.cs.man.ac.uk/tambis) and COHSE [2] (img.cs.man.ac.uk/semweb/cohse)

projects. Specific lines in the workplan are provided for liaison and coordination with standards bodies and other parties. OpenEHR (www.openehr.org) will take the lead in the new CEN TC/251 task force on Electronic Healthcare Records on the basis of the EHCR implantation which is the basis of CLEF.

9. Detailed information about the planned use of Grid technologies and approaches

CLEF is concerned primarily with the “knowledge Grid” and with making the e-Science layer of the Grid available to clinical researchers. The aim of the proposal is to make use of the Grid infrastructure through generic extensions to MyGrid. Much of the Grid architecture and integration work will be done in Manchester, in conjunction with members of the MyGrid team. CLEF will interact directly with low-level Grid protocols, *e.g.* Globus, only if special services are required for Electronic Healthcare Records, *e.g.* enhanced security provisions.

Insofar as practical, CLEF will build on the Metadata services, information extraction and information governance built into the MyGrid architecture and on its mechanisms for “service market places” *e.g.* using UDDI, WSDL and web services. An early effort will be to establish the relationship between the Grid framework and ongoing standards for middleware for Electronic Healthcare Records, *e.g.* 6NET and 6WINIT (see 3.2).

10. Bandwidth requirements and usage (including University provision)

Communication between UCL and Manchester will use SuperJanet for transfers. In the early phases of the project, wide bandwidth transfer will be confined to images to be added to repositories. As the image archiving project matures, transfer rates in the order of 1 Gbps will be required to make co-operative working effective. Integrating bioinformatics databases within CLEF will also require substantial bandwidth. Estimates made for MyGrid suggest that the use of bioinformatics resources will peak at 1 Gbps. SuperJanet is connected to the University of Manchester and UCL campus networks via 1 Gbps links, with an upgrade route to 2.5 Gbps with high speed networks within the campuses. Links to the NHS-Net are of much lower bandwidth although upgrades are under consideration, so all high bandwidth usage will be within the University system.

11. The anticipated end-stage deliverables and their application to human health

CLEF's long run aims are as illustrated in the scenarios in Section 1. In summary they are:

- 1 *Faster, more cost effective clinical trials and experiments linking clinical and genomic information*
- 2 *More effective tailoring of care to individual patients*
- 3 *Improved quality clinical information and reduced workload for clinicians*

In addition, allowing patients greater access to their own records and providing and general information to patients and public are important goals in CLEF's long term programme of research. The information governance, language and metadata techniques developed in this project are an essential steps to achieve those ends.

12. Value for money

CLEF aims to do only those things which are specific to clinical medicine's requirements for e-Science and the Grid. Development of generic e-Science resources will draw on MyGrid and related projects. Electronic Healthcare Records and communication will draw on the established work at UCL and the Royal Marsden. Management of images will draw on the EPSRC/MRC IRC *From Medical Images and Signals to Clinical Information*. At the same time CLEF addresses a problem which is critical for clinical practice as well as clinical e-Science — improved information quality for patient care, quality assurance, and clinical governance.

13. Data security arrangements

A major effort in the project is devoted to information governance and security. An Information Custodian Authority will be constituted to take overall responsibility for confidentiality and security. Within the project, confidentiality will be ensured by routine anonymisation followed by further manual scrutiny of texts for residual identifying information by the participating clinical institutions. In order to avoid any question of unauthorised access to even anonymised data, provision is made for separate servers not connected to the University networks or the Grid. The repositories will only be made available on the Grid when the Information Custodian Authority is satisfied that the security arrangements are adequate and acceptable to relevant legal and clinical ethical authorities.

The analysis of the organisational issues and requirements for governance and security of clinical information will be an important input into determining how Grid security needs to be extended and elaborated for clinical use. Assuming these satisfy the relevant authorities, CLEF will use the Grid security components which exist or are under development in MyGrid — *e.g.* the GSS API within Globus and/or alternative encryption provided by UNICORE from the EUROGrid project.

14. Consideration of IPR issues

The major software and ontological resources in CLEF and MyGrid are open source or freely available in the standards domain. MyGrid has extensive provision for controlling provenance and managing the inclusion of proprietary material with specific restrictions and/or authorisation requirements if required by later collaborations.

15. Consideration of any ethical issues

A major workpackage in the project is devoted to issues of information governance, patient confidentiality, and the ethical issues in the use of patient data in clinical research. This workpackage will take into account, among other resources, the MRC guidelines on *Personal Information in Medical Research*. A special Information Custodian Authority will be constituted to manage issues of access and control. All information used in the project will be fully anonymised and the anonymisation manually scrutinised by the originating institutions. The standards and policies developed will be presented to the relevant ethical committees prior to material being made available.

16. Any information of resources levered from the host institution, other agencies or industry

Manchester Information and Associated Services (MIMAS) a JISC-supported national Grid centre with resource identified to support local E-Science projects including time on: IBM RS6000, (146 PEs), SGI Origin2000, the VR facilities, SGI Origin2000, IBM Beowulf cluster, and Sun E6500. £1m has been set aside within the UCL SRIF investment programme to underpin Grid research and development. Sun Computers and Oracle are providing major discounts on servers and software plus free consultancy. The majority of the effort on organisational issues and clinical resource is being provided independently by the Judge Institute at Cambridge and the Royal Marsden NHS Trust respectively, amounting to several person years of effort each.

References

1. Baker, P, Brass, A, Bechhofer, S, Goble, C, Paton, N, and Stevens, R. *TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview*. in *Sixth International Conference on Intelligent Systems for Molecular Biology, ISMB 98*. 1998. Montreal: pp. 25-34.
2. Bechhofer, S and Goble, CA, *Thesaurus construction through knowledge representation*. Data and Knowledge Engineering, 2001: pp. (in press).
3. de Roure, D, Jennings, N, and Shadbolt, N, *Research Roadmap for e-Science Infrastructure*, . 2001, Comissioned for the EPSRC/DTI Core e-Science Programme (used with permission). pp. 65.
4. Fensel, D, van Harmelen, F, Horrocks, I, McGuinness, D, and Patel-Schneider, P, *OIL: An ontology infrastructure for the semantic web*. IEEE Intelligent Systems, 2001. **16**(2): pp. 38-45.
5. Gaizauskas, R, Cunningham, H, Wilks, Y, Rogers, P, and Humphreys, K. *GATE: An environment to support research and development in natural language engineering*. in *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence*. 1996. Toulouse, France: pp. 58-66.
6. Gaizauskas, R and Humphreys, SK, *Using a semantic network for information extraction*. Journal of Natural Language Engineering, 1997. **3**(2 & 3): pp. 147-169.
7. Hahn, U, Romacker, M, and Schulz, S. *medSynDiKATe: Design considerations for an ontology-based medical text understanding system*. in *AMIA 2000 - Proceedings of the Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care*. 2000. Los Angeles: Hanley and Bulfus: pp. 330-334.
8. Horrocks, I and Patel-Schneider, P. *The generation of DAML+OIL*. in *Description Logics 2001 (DL2001)*. 2001: pp. 30-35.
9. Humphreys, K, Demetriou, G, and Gaizauskas, R, *Bioinformatics applications of information extraction from journal articles*. Journal of Information Science, 2000. **26**(2): pp. 75-85.
10. Krauthammer, M and Hripcsak, G. *A knowledge model for the interpretation and visualisation of discharge summaries*. in *Proceedings of AMIA Fall Symposium 2001*. 2001. Washington DC: Hanley & Belfus: pp. 339-343.
11. NHS National Health Service Executive, *Information for Health: An information strategy for the modern NHS 1998-2005*, . 1998, UK National Health Service Executive.
12. Padkin, A, Rowan, K, and Black, N, *Using high quality clinical databases to complement the results of randomised controlled trials: the case of recombinant human activated protein C*. British Medical Journal, 2001. **323**: pp. 923-926.
13. Power, R, Scot, D, and Evans, R. *What you see is what you meant: direct knowledge editing with natural language feedback*. in *Proceedings of the 13th Biennial European Conference on Artificial Intelligence (ECAI-98)*. 1998: Springer-Verlag: pp. 677-681.
14. Rector, AL, Johnson, PD, T, S, Wroe, C, and Rogers, J. *Interface of inference models with concept and medical record models*. in *Artificial Intelligence in Medicine Europe (AIME)*. 2001. Cascais, Portugal: Springer Verlag: pp. 314-323.
15. Rector, AL, *et al.*, *Reconciling Users' Needs and Formal Requirements: Issues in developing a Re-Usable Ontology for Medicine*. IEEE Transactions on Information Technology in BioMedicine, 1999. **2**(4): pp. 229-242.
16. van Deemter, K and Power, R. *Authoring multimedia documents using WYSIWYM editing*. in *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. 2000. Saarbruecken, Germany: pp. 222-228.