A Co-operative Clinical E-Science Framework (CLEF): Joining up Healthcare and Clinical Research

Jeremy Rogers[1] Adel Taweel[1] Alan Rector[1] Dipak Kalra[2] David Ingram[2] Jo Milan[3] Peter Singleton[4] Robert Gaizauskas[5] Mark Hepple[5] Donia Scott[6] Richard Power[6]

[1]BioHealth Informatics Forum, Department of Computer Science, University of Manchester
[2]Centre for Health Informatics & Multiprofessional Education, University College London
[3]Royal Marsden Hospital NHS Trust
[4]Judge Institute, University of Cambridge
[5]Department of Computer Science, University of Sheffield
[6]Information Technology Research Institute, University of Brighton

jeremy@cs.man.ac.uk www.clinical-escience.org

Abstract

The CLEF project aims to establish a secure socio-technical framework that enables sharing patient data for the purposes of research whilst maintaining patient privacy and confidentiality. The value of shared data is increased by integrating, within a secure repository, both existing structured information (lab reports etc) with information extracted from texts (clinic letters), and using clinical inferencing and filtering techniques to derive a canonical view of the record called the 'chronicle'. Statistical disclosure control and Language generation technologies are used to simplify and control access to this complex resource.

Introduction

Our increasing ability to gather information at the molecular level (genomic, proteomic etc) contrasts with our inability to gather information about patients. Clinical medicine and post-genomic research need better patient data on the progress of disease or their response to treatment, to answer the questions:

What happened and why?
What was done and why?

The Problem

Today, these apparently simple questions can only be answered by manually examining each patients' notes – a time consuming process whether the notes are electronic or paper. It therefore remains very difficult either to systematically measure the quality of clinical care across even modest patient populations, or to properly investigate the genetic factors that influence the course of a disease in an individual or their response to any treatments.

The CLEF Approach

CLEF focuses on those specific technologies needed to permit, and enable, better clinical information capture, integration and sharing:

- *Privacy and security policy and requirements to protect patients*
- *Information extraction from multiple texts to acquire the information*
- *Language generation to support easy querying and presentation of information*
- *Integration and 'chronicalisation' of clinical information*
- *Knowledge Sources to recognise implied events and their interrelationships*
- *Standards for data and metadata*

CLEF and e-Science

CLEF depends on other e-Science projects for:

- *Grid based security including role based authorisation to implement the privacy and security policies*
- *Workflow, provenance, and web/grid service architectures and registries*
- *Semantic Web/Grid tools and technologies to support the chronicle.*

The CLEF Data Cycle

Figure 1 illustrates two interlocking data cycles:

Left hand cycle: data resulting from normal recording of clinical activity is anonymised and depersonalised before analysis, integration and

summarisation. Individual patient summaries can be fed back to clinicians, under strict re-identification controls.

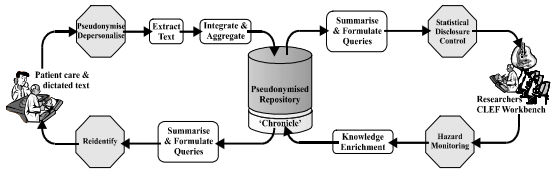

Figure 1: Basic CLEF Information Flow

Right hand cycle: the repository of anonymised clinical information can be queried by researchers with appropriate authorisation.

Privacy and Security

Clinical data is extremely sensitive. Its use outside the clinical process is tightly controlled. Correspondingly, a large part of the CLEF infrastructure concerns the privacy and security of patient data. It recognises that no technical solution can be perfect in this respect.

The single most critical criterion for convincing clinical research committees to permit sharing of clinical data through any eScience framework is demonstrating that the benefits of sharing outweigh the quantified risks, and that organisational measures are in place both to monitor and further minimise that risk.

Overall, a socio-technical approach is required, but CLEF's ability to use the wider eScience technical infrastructure is contingent on successfully reconciling existing healthcare standards with eScience standards, particularly those relating to security, confidentiality and accountability.

Information Extraction

Much important clinical information is contained only in unstructured text, and CLEF anticipates that this will continue for the foreseeable future. CLEF is adapting and evaluating information extraction technologies, seeking to exploit special features of the clinical and cancer domains. These include in particular the highly controlled sublanguage, and the repetition of important information across multiple documents [1].

The CLEF Chronicle

The classic problem for electronic health records is to maintain a faithful, secure, non-repudiatable record of what healthcare workers have heard, seen thought and done [2]. The CLEF Electronic Health Record repositories follows standards designed to achieve these aims – *e.g.* OpenEHR [3], CEN standard 13606, and associated development of "archetypes"[4].

However, the central issue for CLEF is different – to infer a single coherent view of each patients' history from the myriad documents and data in the Electronic Health Record, and to align this with other similar patients in aggregate for querying and research.

CLEF does not restrict itself only to the literal information of what was said in the documents. What was unsaid but only implied or otherwise obvious to the human reader are also important. If a bone scan report states "only osteoporotic changes", CLEF must recognise that this phrase also indicates that there were "no bony metastases found". Similarly, it is not enough to know only that a patient prematurely discontinued chemotherapy; what side effect or intercurrent illness intervened?

Figure 2 shows a representation of a patient record as a graphical timeline, developed in the course of CLEF requirements gathering. A clinician can quickly infer the connections between the events displayed from their temporal juxtaposition – for example that the first episode of radiotherapy was to treat the first relapse. An effective computer based 'chronicle' must make those inferences explicit.

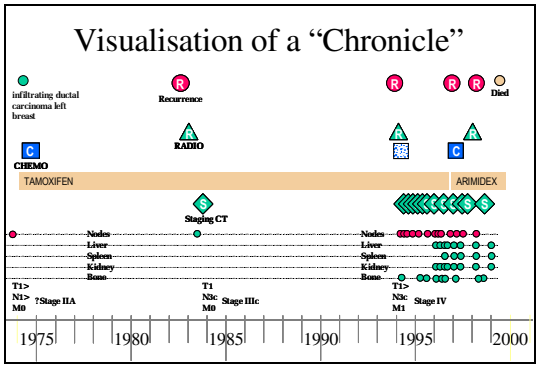

Figure 2: Visualisation of a 'Chronicle'

At the heart of CLEF, therefore, is the compilation of a single coherent and non-redundant "chronicle" for each patient from the distributed heterogeneous, and often repeated, information that makes up the traditional medical record.

The chronicle therefore draws on structured information (laboratory reports etc.) from the traditional record together with the output of information extraction on the narrative element (clinic letters, investigation reports). These data

are integrated, filtered to remove duplication and redundancy, cross-referenced and then augmented to insert new information that was not originally explicitly present in the source material, for example translating a series of low blood results recorded at distinct time points into an entry stating that the patient was anaemic for the duration.

The chronicle represents the medical record not as a series of disjoint electronic documents but as a single semantic network of nodes and relations (Figure 3). Each node and relation has provenance links to those original documents and inference rules that supports their existence in the chronicle.

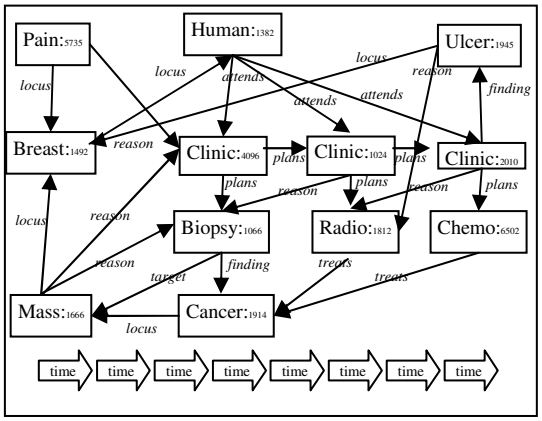

Figure 3: Chronicle structure

Sets of chronicles then become data structures that can be easily aligned on "index events" – diagnosis, first treatment, relapse, etc.- and aggregated for statistical analysis to answer questions such as:

- Of patients with breast cancer and with a particular genetic profile, compare the time to first recurrence for those treated with Tamoxifen as against those treated with a new agent in clinical trial
- How many dropped out of each treatment and why?
- How many required supplementary therapy for the side effects of treatment and why?
- Compare the survival of clinical trial patients who interrupted their radiotherapy to take a vacation, with those who did not.

Query Formulations

For the data in the CLEF repository to be useful to researchers in the right hand data cycle (Figure 1), it must be easy to interrogate. A variety of graphical and textual query interfaces are being explored, but the primary interface is being designed around techniques from language generation known as 'What you see is what you meant' [5]. An example is given in Figure 4, showing a partly specified query (above) and the reply to a different and more completely specified query (below).

Discussion

CLEF has permission to work with data concerning deceased patients, and already has a repository populated with more than 300k text documents relating to twenty thousand deceased patients. Further data, particularly the structured laboratory and prescribing records, will be integrated by the end of 2004. Work at national policy level is underway to identify a path that would ultimately allow work with data from living patients.

Query
Treatment profiles: Patients who received **[this type of treatment],** compared with patients who **[did not]**.
Outcome: Percentage of patients alive after **[5 years]**.
Relevant subjects: Patients with **[cancer]** of the **[pancreas]**
Answer|
It was found that out of 1790 patients diagnosed with cancer of the pancreas, 1300 had a pancreaticoduodenectomy and 490 didn't. Out of the 1300 patients who had a pancreaticoduodenectomy, 890 (68.46%) were alive after 5 years. Out of the 490 patients who did not have a pancreaticoduodenectomy, 87 (17.75%) were alive after 5 years.

Figure 4: WYSIWYM Example

Preliminary information extraction experiments are underway, and a prototype query workbench and chronicle generator have been constructed.

However, in addition to technical solutions, CLEF must also address significant organisational issues. There is considerable added value to UK clinical research if the open and distributed computing ethos of eScience can be interfaced with the closely regulated and centralised requirements of the NHS, but achieving this interface is not straightforward, particularly while both programmes are moving forward independently with very different priorities and timescales.

Acknowledgements

CLEF is supported in part by grant G0100852 from the MRC under the e-Science Initiative. Special thanks to its clinical collaborators at the Royal Marsden and Royal Free hospitals, to colleagues at the National Cancer Research Institute (NCRI) and NTRAC and to its industrial collaborators – see www.clinicalescience.org.

References

1. Gaizauskas, R., Hepple, M., Davis, N., Guo, Y., Harkema, H., Roberts, A. and Roberts, I., AMBIT: Acquiring Medical and Biological Information from Text. in *Second UK E-Science "All Hands Meeting"*, (Nottingham, 2003), (in press)
2. Rector, A., Nowlan, W. and Kay, S. Foundations for an Electronic Medical Record. *Methods of Information in Medicine*, *30*. 179-186.
3. Ingram, D. GEHR: The Good European Health Record. in Laires, M., Ladeira, M. and Christensen, J. eds. *Health in the New Communications Age*, IOS Press, Amsterdam, 1995, 66-74
4. Beale, T., Archetypes: Constraint-based domain models for future-proof information systems. in *OOPSLA-2002 Workshop on behavioural semantics*, (available from http://www.oceaninformatics.biz/publications/archetypes_new.pdf, 2002).
5. 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), Spring-Verlag, 677-681.