CLEF – Joining up Healthcare with Clinical and Post-Genomic Research

Alan Rector* Jeremy Rogers* Adel Taweel* David Ingram Dipak Kalra** Jo Milan*** Peter Singleton‡ Robert Gaizauskas† Mark Hepple† Donia Scott†† Richard Power††**

*BioHealth Informatics Forum, Department of Computer Science, University of Manchester
**Centre for Health Informatics & Multiprofessional Education, University College London
***Royal Marsden Hospital NHS Trust
‡Judge Institute, University of Cambridge
† Department of Computer Science, University of Sheffield
††Information Technology Research Institute, University of Brighton

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

ABSTRACT

CLEF aims to join up clinical care and biomedical research. It is developing methods for managing and using pseudonymised repositories of the long-term patient histories which can be linked to genetic, genomic and image information or used to support patient care. CLEF concentrates on removing key barriers to managing such repositories – ethical issues, information capture, integration of disparate sources into coherent "chronicles" of events, user-oriented mechanisms for querying and displaying the information, and compiling the required knowledge resources. This paper describes the overall information flow and technical approach designed to meet these aims within a Grid framework. Details of work on language technology and ethical issues are discussed in separate papers at this conference.

Background: Common need for improved clinical information

Information on the long-term course of patients' illnesses and treatments is needed both to improve clinical care and to enable post genomic research. Scientists and clinicians alike must answer the fundamental questions about patients' histories:

What was done and why?
What happened and why?

Simple as these questions sound, they are difficult to answer using current technology without recourse to manual examination of patients' notes – a time consuming and expensive process. Yet, without answers to these questions, it is difficult either to measure the quality of care or to investigate the factors affecting onset of disease or outcome of care.

Hence CLEF aims to provide a repository of well organised clinical histories and which can be queried and summarised both for biomedical research and clinical care.

Barriers to improved clinical information

CLEF addresses key barriers to capturing clinical information and managing it in such repositories.

- *Privacy, consent, and security* – at all levels: policy, organisational structure, and technical implementation. [4]
- *Information capture* – much of the information required is available only as dictated texts from which it must be extracted.
- *Information integration and 'chronicalisation'* – in their raw form, clinical records consist of hundreds of diverse documents. To be useful, a coherent 'chronicle' of events must be inferred from them.
- *Information presentation and summarisation* – The CLEF repository is to be used by clinicians and scientists, not IT specialists. The questions to be asked are complex and may require information from many sources. The questions to be asked can be only partly predicted in advance.
- *Knowledge resources* – all of the above tasks are knowledge intensive. They require recognising which information is significant in which context and recognising the implied meaning of information. While many knowledge resources exist, many do not, and coordinating those that do is a task in itself
- *Standards* – cooperation requires standards, which are only just emerging.

The basic CLEF Information Flow

The basic CLEF information flow is shown in Figure 1. The top cycle represents the information capture from the clinician into the repository and its use either for clinical care of research.

Starting with the "Patient care and dictated text" at the top of the diagram, the flow is:

- *Capture* of the information. Some information comes from dictated and transcribed text .Other information comes directly from hospital information systems – *e.g.* laboratory results, prescriptions, etc.
- *Pseudonymisation* of all information by removal of overt identifying items – name, date of birth, etc - and by providing a CLEF Entry identifier that can only be reversed by the NHS provider or a trusted third party.
- *Depersonalisation* of the texts to remove any residual information that might risk identification – *e.g.* names of relatives, nick names, place names, unusual occupations, etc.
- *Information extraction* of key information from the texts into predefined "templates", possibly with the help of context provided by the information already in the repository.
- *Integration into the health record repository* of all information including laboratories, radiology, and potentially genomic analyses
- *Constructing the chronicle* to infer a coherent view of the patient's history. Typically the same information occurs in many different documents with different levels of granularity, clarity and sometimes conflicts that must be reconciled.

From this point the information can go in two directions.

- Back to the clinicians in the form of summaries for patient which can be re-identified by the hospital. Providing a concise up-to-date summary of a patients' condition is a prime request of clinicians for improving patient care.
- On to researchers in response to queries.

The overall use of the repository for research is under the control of an Ethical Oversight Committee to which researchers must apply to gain access to anything except pre-computed results and metadata. Despite other precautions, it is assumed that if individual records can be read in detail, there is the risk of identification. Therefore, all information is treated with the as if it were for identifiable. [4].

Most researchers will be accredited only for aggregated results controlled through the privacy enhancing technologies. Special permission of the Ethical Oversight Committee is required to gain access to individual patient records, since there is always the risk that an individual patient can be recognised from the course of their illness.

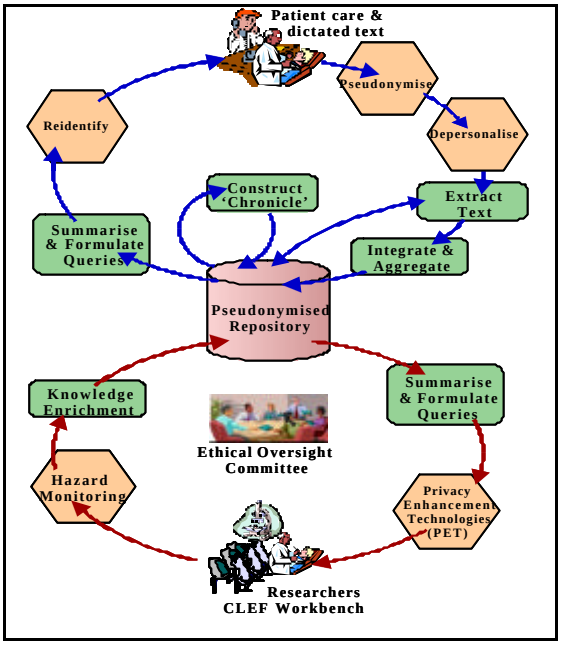

Figure 1: Basic CLEF Information Flow

Privacy enhancing technologies are used to monitor the query process to avoid the risk that patients can be re-identified on the basis of other information known about them as described in [6]. All researchers have a duty of care to be aware of potential hazards and risks of compromising patient confidentiality that must be reported along with results.

Underlying Technologies

Language Technology

Information Extraction from multiple texts

The key characteristic of cancer patients for information extraction is that they are seen repeatedly, so that there are usually many texts describing the same events from different points in time. Typical lifetime records consist of 100-200 text documents plus hundred or thousands of items of structured data derived from laboratory, pharmacy or other hospital subsystems. Even a single episode of care is likely to generate several texts. The precision of information extracted can therefore be built up gradually as increasing numbers of documents from one patient are processed. For further details see [2].

Depersonalisation and the use of dead patients' records for analysis

A second function for language technology is to "depersonalise" texts – *i.e.* to remove comments

in the text which might accidentally make it likely that the individual could be identified – such as nick-names, place names, employers, etc. This is a relatively standard task for language technology using "named entity recognition" [7]. Until the methods are proven to the satisfaction of the oversight committee, CLEF is confining itself to records of patients who are deceased.

Information Integration into standard healthcare record formats

The past decade has seen extensive work on structuring healthcare records for patient care. CLEF draws particularly on the work on OpenEHR which forms the basis for the new CEN standard for information interchange [3]. The structure of information in the healthcare record is built out of standard "archetypes". "Archetypes" are reusable elements which facilitate interoperability and re-use and which can evolve as medical knowledge and practice evolves. The notion of archetype has also been taken up by HL7 [3], the major standardisation body for healthcare informatics, and is closely related to the notion of "elements" in the US National Cancer Institutes CaCore Architecture[1].

The fundamental requirement on the medical record is that it is a faithful log of what has happened and who takes responsibility for actions taken. Information can never be deleted, and although all information is time stamped there is little notion of duration or relative time.

There is no attempt in the medical record *per se* to provide a single coherent view of the patient, rather its purpose is to record the different views that have been taken about the patient at different times by different carers. Finally, the medical record emphasises *what* rather than *why*.

"Chronicalisation"

By contrast the CLEF chronicle is an attempt to form a coherent view of the best inference about what has actually been done and why; what has happened and why. Time relative to 'index events' and durations are at least as important as calendar time. In general, this is an extremely difficult transformation. Fortunately in the cancer domain, it appears achievable, albeit with

A major advantage of working in cancer domain is that patients' histories are marked by relatively discrete events – diagnosis, definitive treatment, recurrence, death, etc. Furthermore, most index events are described repeatedly in varying detail. This is particularly important when dealing with records from a referral hospital where the system usually will start in the "middle of the story". For example, first document might simply mention breast cancer in the past, concentrating on the current recurrence. A summary later might give a date for a mastectomy but no details of the tumour type. Eventually, perhaps after information from the referring hospital was received, a definitive statement of the time, tumour, spread, and treatment might be found. Subsequent notes might again refer to the initial cancer vaguely while concentrating on current concerns. Thus the overall 'chronicle' comes into focus gradually as information accumulates.

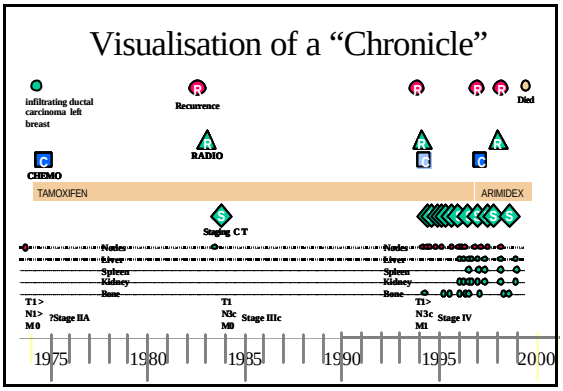

Figure 2: A patient chronicle in graphical form

Furthermore, reasons for actions are rarely given explicitly. That Tamoxifen[2] was given to treat the breast cancer and Paracetamol for pain is taken as too obvious to be mentioned. That chemotherapy was suspended when the patient complained of severe nausea may be stated, but the causal link is merely implied. Constructing the "why" part of the chronicle is therefore a knowledge intensive task. The certainty of the inferred constructs varies, so the evidence and certainty need to be recorded.

Query formulation, WYSIWYM, and Language Generation

For the data in the Electronic Patient Repository to be useful, it must be easily accessible to scientists and clinicians. The interface to the repository of health records and chronicles is being designed around techniques from language generation known as WYSIWYM –"What you see is what you meant"[1, 5] supplemented by various visual or graphical presentations. An example is given in Figure 3. The next stage of the project will include user studies to ensure that the interface meets users' priorities.

[1] http://ncicb.nci.nih.gov/core

[2] An anti-oestrogen used to treat breast cancer and for no other purpose

Query

Treatment profiles: Patients who received **[this type of treatment],** compared with patients who did not. *Outcome measure*: Percentage of patients alive after **[this interval of time]**.

Relevant subjects: Patients with **[this type of cancer]**

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 3: Example of WYSIWYM query formulation and natural language response

Knowledge resources required

All the key technologies in CLEF are knowledge intensive. The overall approach in CLEF is based on "ontology anchored knowledge bases"– knowledge bases anchored in common conceptual models but conveying additional domain knowledge about the concepts represented. Examples include which drugs are used for which purposes, the significance of different results from different studies, the fact that a seemingly positive finding such as "evidence only of degenerative changes" may in practice convey the negative information that "no metastases were found". Some of this information exists in established resources such as the UMLS[3]. However, much of it needs to be compiled. CLEF works with both myGrid[4] and the new CO-ODE[5] project to developer-usable knowledge resources and tools.

Metadata in the Repository

The CLEF repository requires at least four types of metadata:

- Resource information: what is in the repository so that it can be found
- Provenance information: where information comes from
- Usage and workflow information: how the information has been used, including information allowing monitoring potential compromises of privacy
- Annotations on certainty and evidence: what inferences have been made on the basis of what evidence with what confidence

The first three appear analogous to metadata within myGrid and related projects. The fourth is more specific to CLEF. Metadata standards also need to take into account emerging standards for annotating clinical trials and other areas of biomedicine.

[3] http://umlsks5.nlm.nih.gov
[4] myGrid.semanticweb.org
[5] www.co-ode.org

Discussion

CLEF is aiming to demonstrate a broad integration of clinical information joining up patient care with basic biomedical research. It builds on the basis of myGrid and other E-Science projects to bring their insights to the clinical domain. It is complementary with projects such as the National Cancer Tissue Resource and National Translational Cancer Network that focus more on actual specimens and genomic information *per se*. It seeks to provide clinical and knowledge resources that will be re-usable, for example within the broad framework being overseen by the National Cancer Research Institute (NCRI) and to lessen the barriers to using clinical information in collaborative research.

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.clinical-escience.org.

References

1. Bouayad-Agha, N., Scott, D. and Power, R. Integrating content and style in documents: a case study of patient information leaflets. *Information Design Journal, 9* (2-3), 161-176.
2. 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).
3. Kalra, D., Austin, A., O'Connor, A., Patterson, D., Lloyd, D. and Ingram, D., Design and Implementation of a Federated Health Record Server, in *Toward an Electronic Health Record Europe 2001*, (2001), Medical Records Institute for the Centre for Advancement of Electronic Records Ltd., 1-13.
4. Kalra, D., Singleton, P., Ingram, D., Milan, J., MacKay, J., Detmer, D. and Rector, A., Security and confidentiality approach for the Clinical E-Science Framework (CLEF), in *Second UK E-Science "All Hands Meeting"*, (Nottingham, UK, 2003), (in press).
5. Power, R., Scot, D. and Evans, R., What you see is what you meant: direct knowledge editing with natural language feedback, in *Proc ECAI-98)*, (1998), Spring-Verlag, 677-681.
6. Sweeney, L. k-anonymity: a model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10* (5), 557-570.
7. Wakao, T., Gaizauskas, R. and Wilks, Y., Evaluation of an Algorithm for the Recognition and Classification of Proper Names, in *Proc of the 16th Intl Conf on Computational Linguistics (COLING96)*, (Copenhagen, 1996), 418-423