



# THE WHITE ROSE GRID

## e-Science Centre of Excellence

## CLEF: CLinical Escience Framework

### Introduction

CLEF aims to develop a high quality, secure and interoperable information repository, derived from operational electronic patient records to enable ethical and user-friendly access to patient information in support of clinical care and biomedical research.

High quality, integrated clinical information is at the intersection of clinical research, evidence-based health care and the clinical application of genetic and genomic research. A coherent clinical information framework is required to meet the needs of patients, their families and carers, clinical professionals and biomedical scientists, health care enterprises and the public at large.

typed; alternatively they are laboriously coded or annotated manually, usually in incompatible formats that lack rigour and hence cannot be scaled up or aggregated effectively.

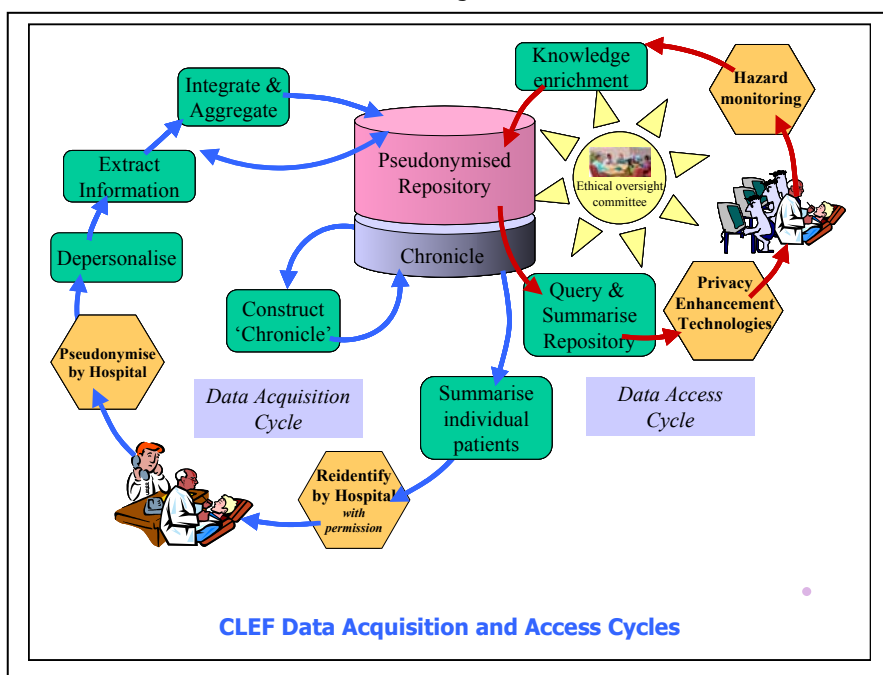


We face an escalating imbalance between the richness of our ability to collect very large sets of genomic, image, and other technical information, and the poverty of our means to describe the scientific and clinical significance of that information.

This same inability to deal effectively with clinical information is a key limitation to our using informatics to support safe, evidence-based healthcare and to gather the information needed to deliver clinical governance and other strategic goals of the NHS.

The Natural Language Processing Group at the University of Sheffield is addressing the information capture and integration barrier by developing tools that automatically extract key medical information from textual clinical notes. The extracted information is added to the central repository in a structured format, where it can be queried and combined with other information already in the repository into a comprehensive summary of patients' treatments and conditions over time.

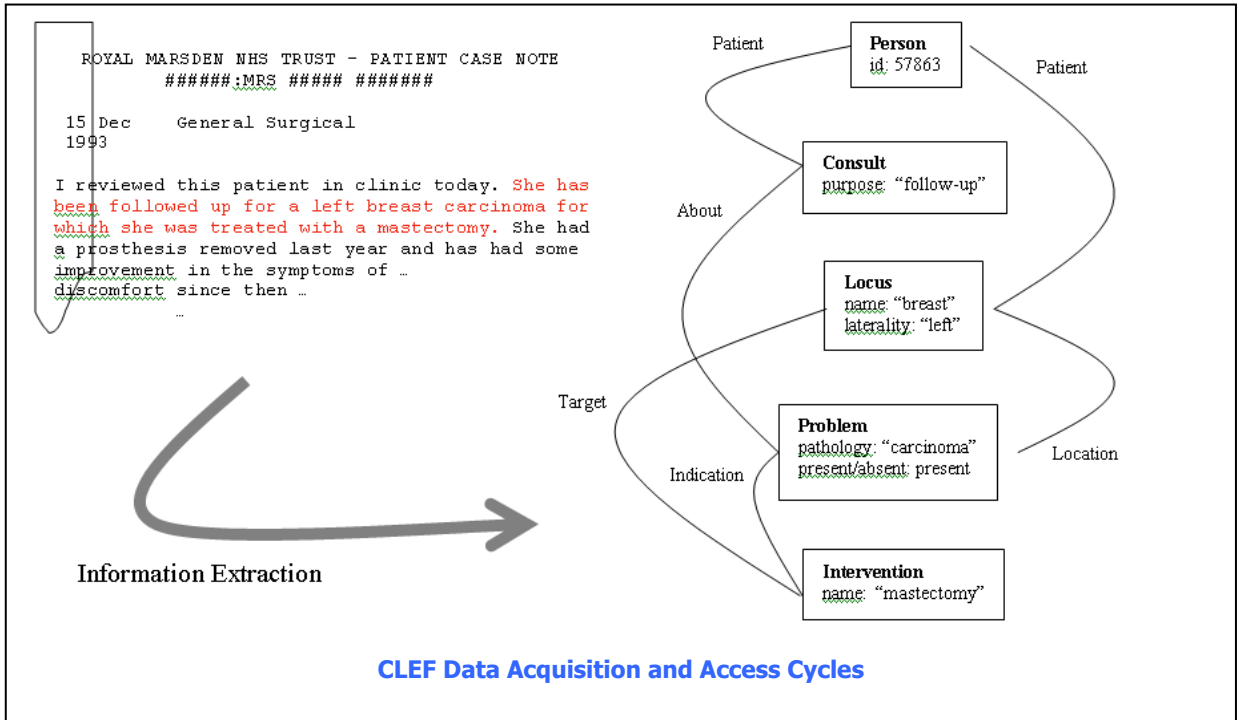
The Information Extraction process comprises three major stages: lexical and terminological processing, syntactic and semantic processing and discourse processing. The first stage identifies and classifies relevant entities that occur in a clinical text, using a general, large-scale terminology resource tuned to recognizing medical terminology. Medical terms that are recognized include drugs, problems (i.e., symptoms and diseases), anatomical



CLEF Data Acquisition and Access Cycles

Capture, integration and presentation of descriptive information is a major barrier to achieving such a framework. Clinical histories, radiology and pathology reports, annotations on genomic and image databases, technical literature and Web-based resources all typically originate as text. Often they are dictated and then





structures, and investigations and interventions.

The second stage produces a (partial) syntactic and semantic analysis for each sentence in the text. The third step integrates these analyses into a discourse model which represents the semantic content of the text. This step involves the application of rules which recognize relationships between entities expressed within and across sentences, e.g., that an investigation has indicated a particular condition, which, in turn, has been treated with a particular intervention. The information to be extracted is then read off from the discourse model and stored in structured objects which are imported into the information repository.

CLEF is a collaborative project between the University of Manchester, University College London, Royal Marsden Hospital, University of Cambridge, University of Sheffield and Open University.

Members of the CLEF team at the University of Sheffield are: Prof. Rob Gaizauskas, Dr. Henk Harkema, Dr. Mark Hepple, Angus Roberts, Ian Roberts, Mark Tice, and Dr. Andrea Setzer. For further information about the CLEF, please visit

<http://nlp.shef.ac.uk/clef/> and <http://www.clef-user.com/>.

CLEF is funded by the UK Medical Research Council, research grant number MRC 60086. CLEF is coming to an end in 2005 but the project has received follow-on funding to roll out the framework developed to hospitals and clinical services.

**Further Information**

Contact:



Rob Gaizauskas (email: [R.Gaizauskas@sheffield.ac.uk](mailto:R.Gaizauskas@sheffield.ac.uk) )

The Project Web site:

<http://nlp.shef.ac.uk/clef/>

**We face an escalating imbalance between the richness of our ability to collect very large sets of genomic, image, and other technical information, and the poverty of our means to describe that information.**



The University of Sheffield.



THE UNIVERSITY of York

