MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

Text Mining: beyond the CAQDAS?

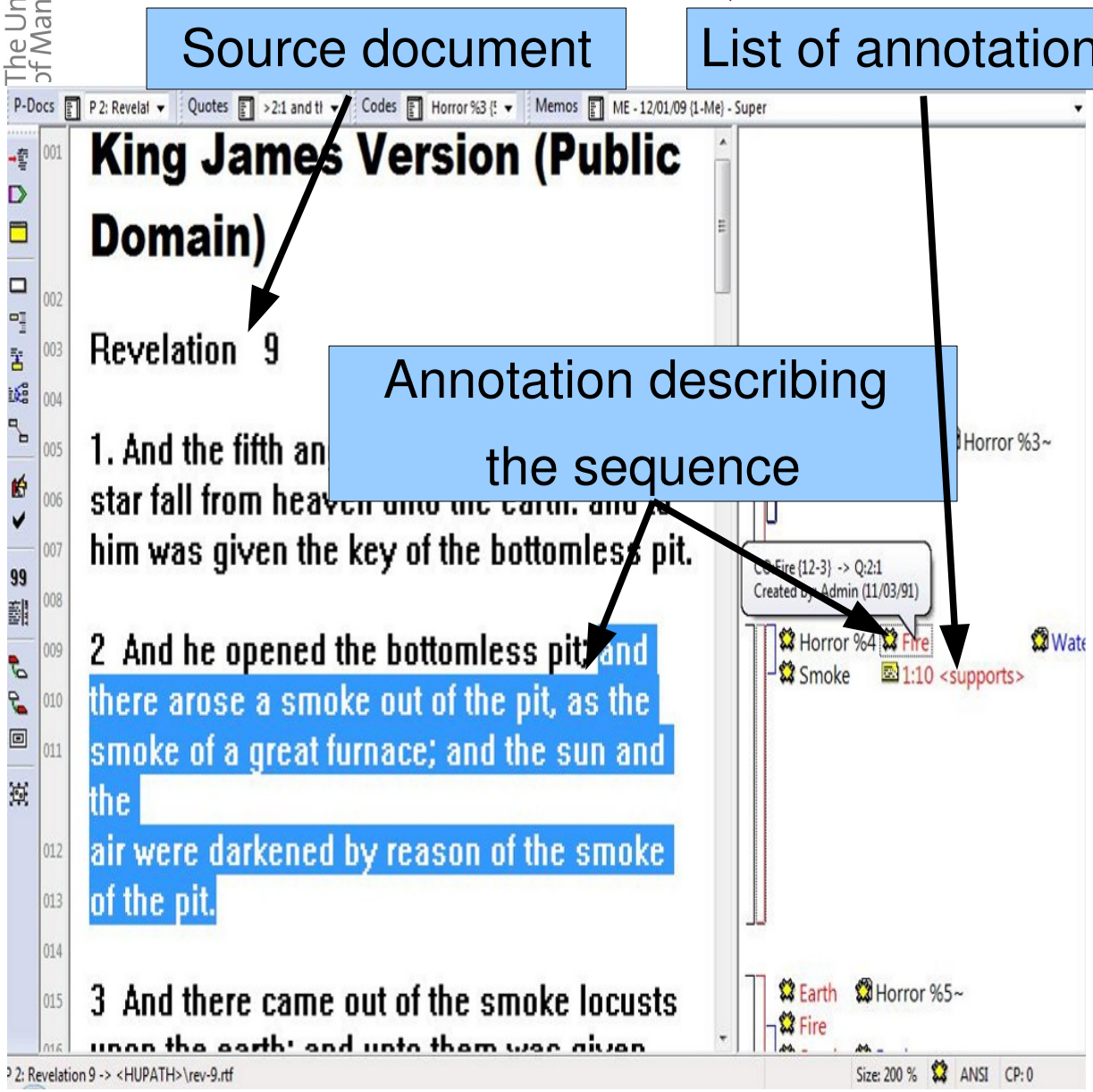
Davy Weissenbacher, Brian Rea,

Sophia Ananiadou

National Centre of Text Mining

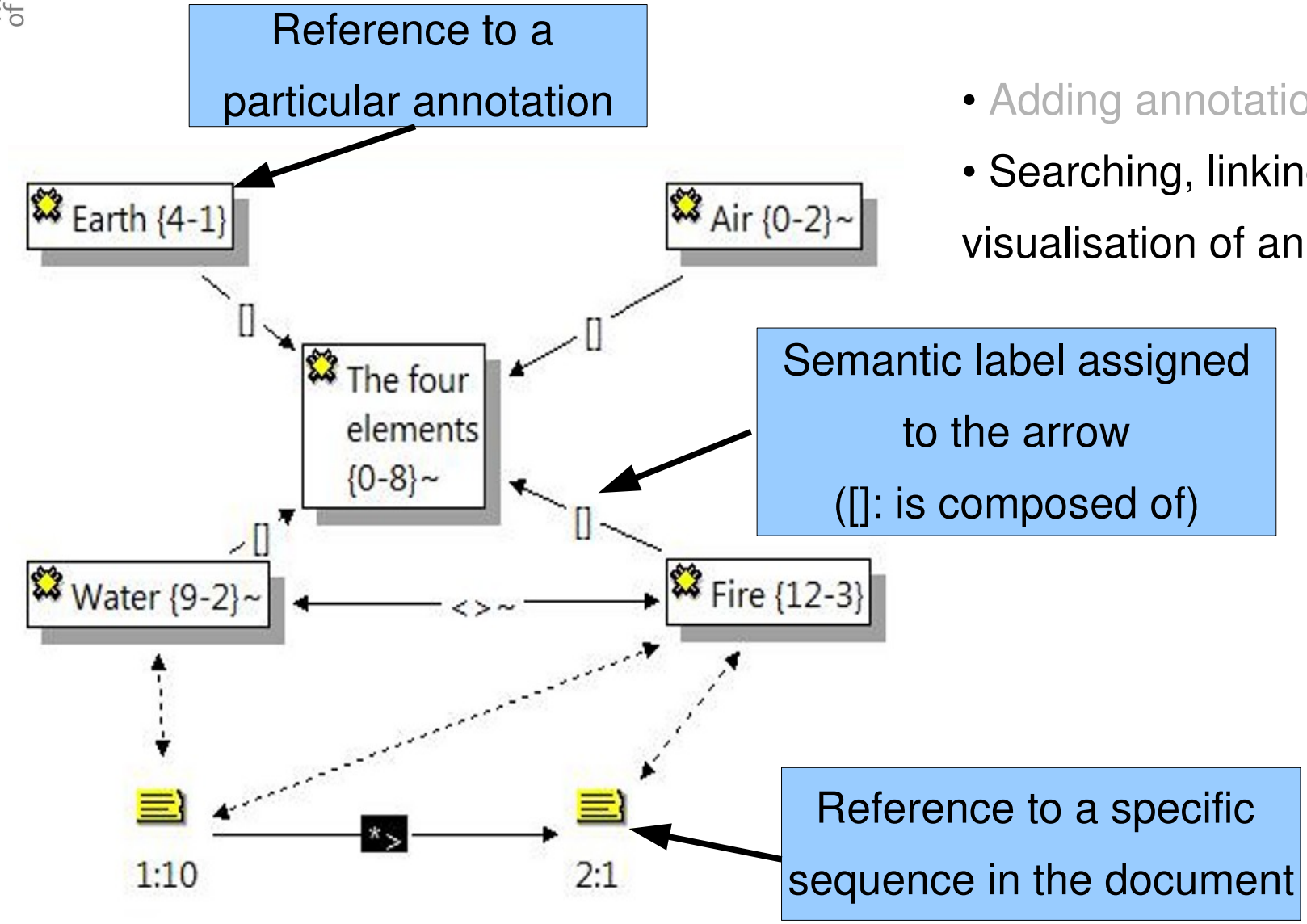
{firstname.surname}@manchester.ac.uk

What a CAQDAS software do?



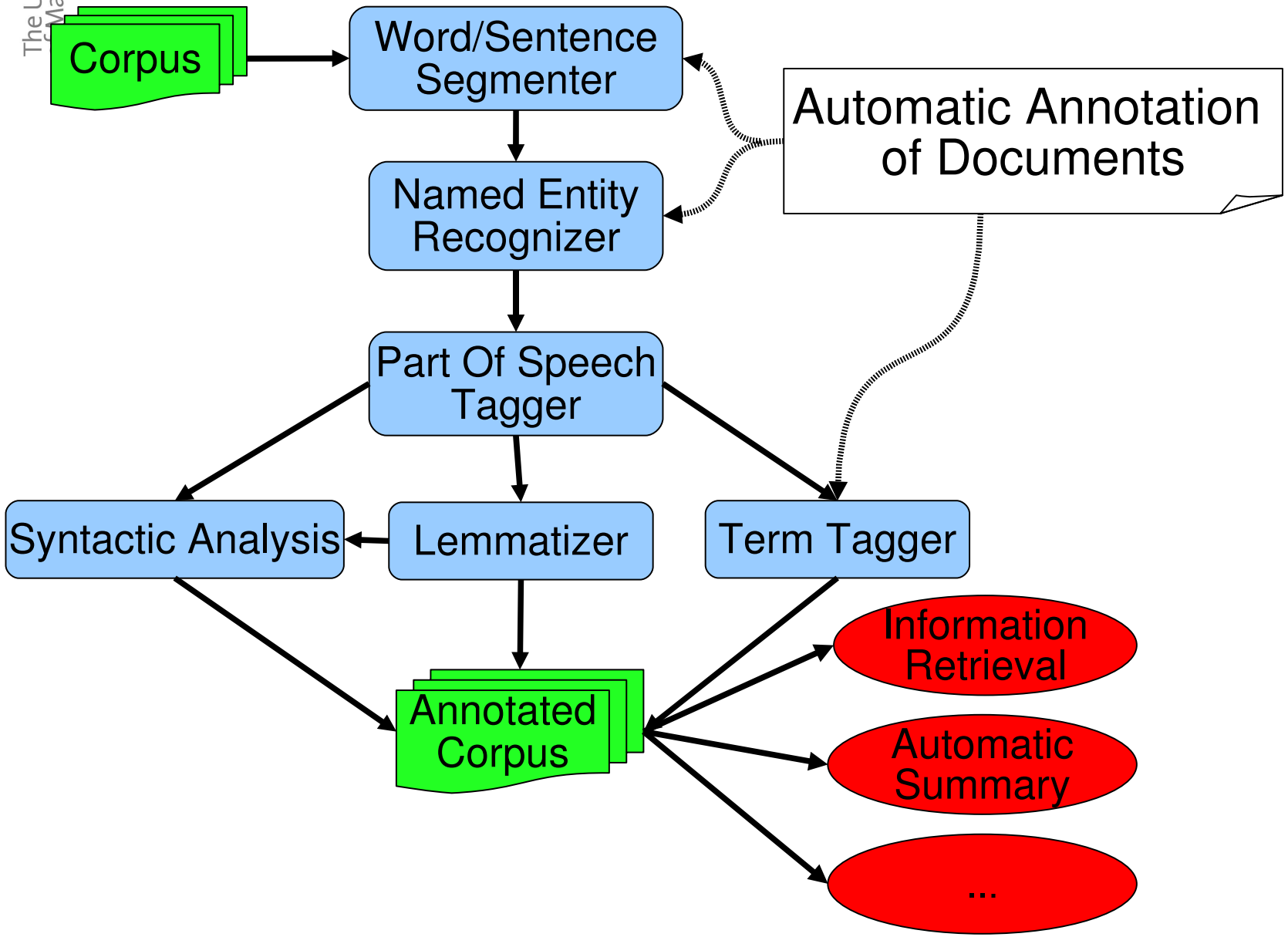
- Adding annotations
- Searching, linking and visualisation of annotations

What a CAQDAS software do?



- Adding annotation
- Searching, linking and visualisation of annotations

What is a Text Mining (TM) Software?



MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

Are the CAQDAS and TM software competitors?

- CAQDAS and TM software are designed to add annotations but:
  - CAQDAS: human annotation (Hundreds of documents)
    TM : automatic annotation (Millions of documents)
  - CAQDAS: Semantic and Pragmatic annotations
    TM : Syntactic and Simple semantic annotations

How can TM techniques complement CAQDAS software?

- TM techniques enrich CAQDAS:
  - QDA Miner + Wordstat: stoplist for word frequency, lemmatizer, thesaurus for retrieving sequence to annotate, clustering of documents
  - Qualrus: machine learning techniques to propose sequences to annotate
- TM techniques are used to:
  - Extend the user queries
  - Focus the user attention on the pertinent sequences

➔ The ASSIST Project: evaluate the benefits of TM for frame analysis of Media

MANCHESTER 1824

The University of Manchester

NaCTeM
The National Centre for Text Mining

ASSIST project

- Aims to deliver a service for searching and qualitatively analysing social science documents
- NaCTeM is designing and evaluating an innovative search engine embedding text mining components
  - **Domain knowledge** facilitates expansion of user queries
  - Real Time **clustering** of search results
  - **Term extraction** for improved browsing capabilities
  - **Semantic Information enrichment** for targeting the main topics
- Final deliverable will include a web demonstrator for further integration into JISC e-Infrastructure
- NaCTeM local project website: http://www.nactem.ac.uk/assist/

Technical Characteristics

Multi-format documents

Conversion tools

.PDF with *pdfbox*

.DOC with *POI*

.HTML with *Jtidy*

.XML

TM components

•Named Entity Recognizer

BaLIE

•Term Extractor

Termine

•Anaphora resolver

Bayaphora

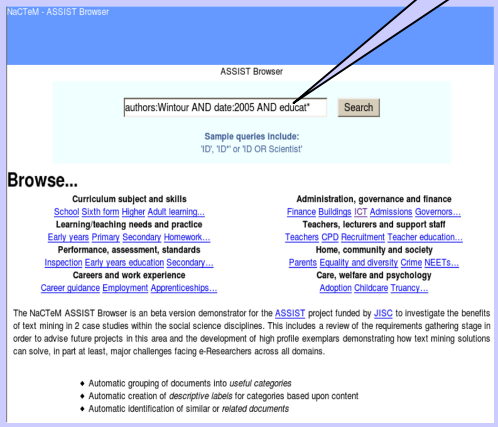
•Lexical Chain extractor

Search Engine

Lucene

User Query

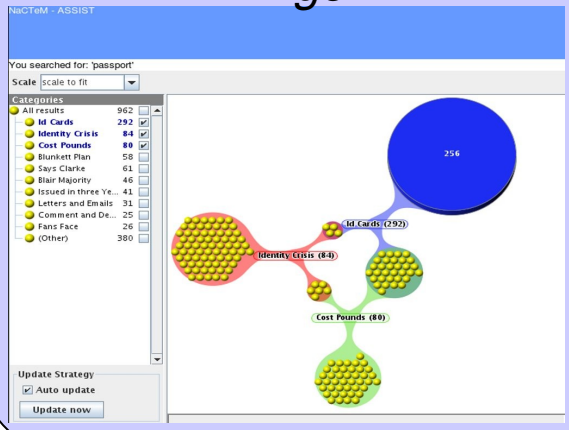
Web Query Interface



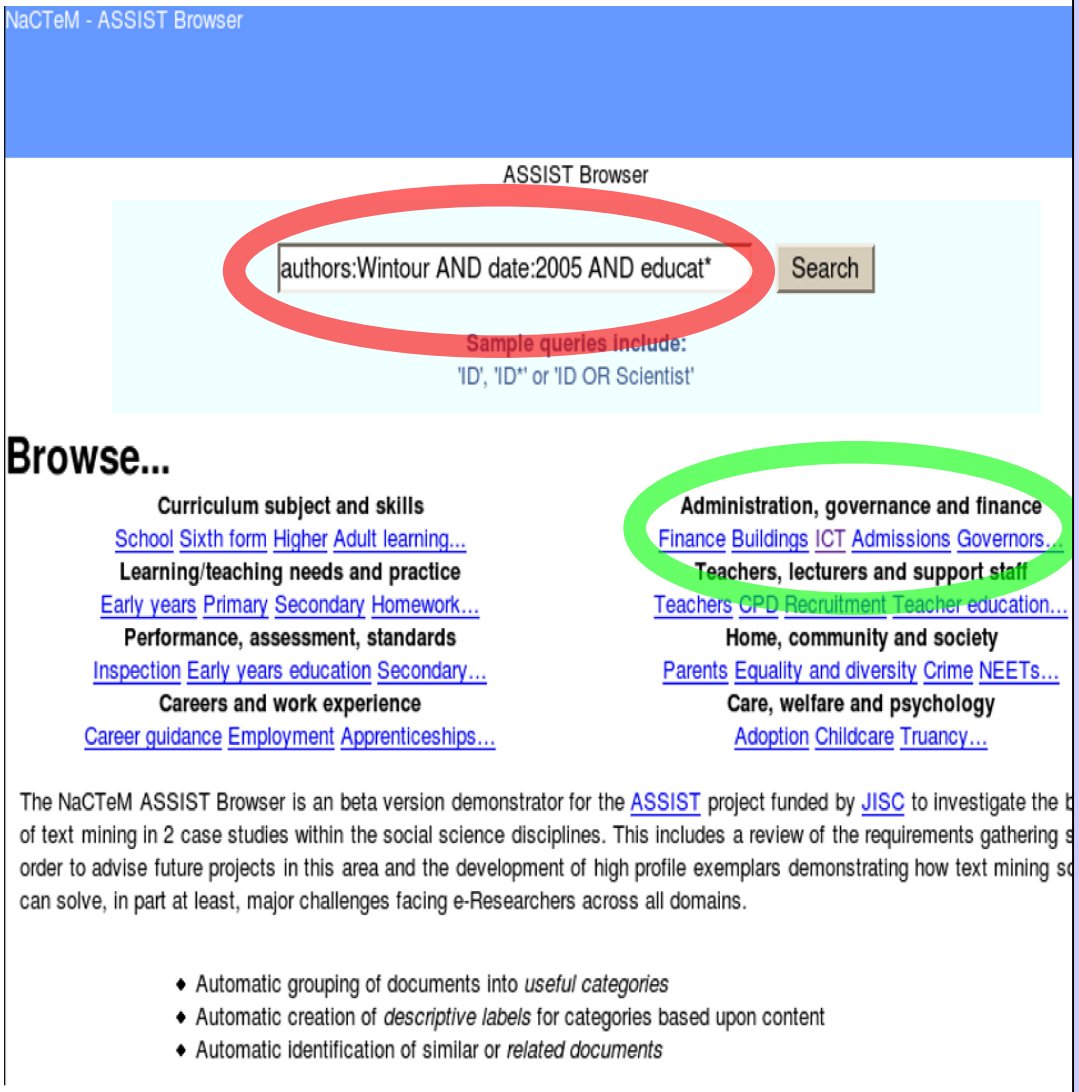
Indexed Documents

Search result clustering

Lingo



Query interface



Expanding the standard query interface

✓ Semantic operators to build complex queries

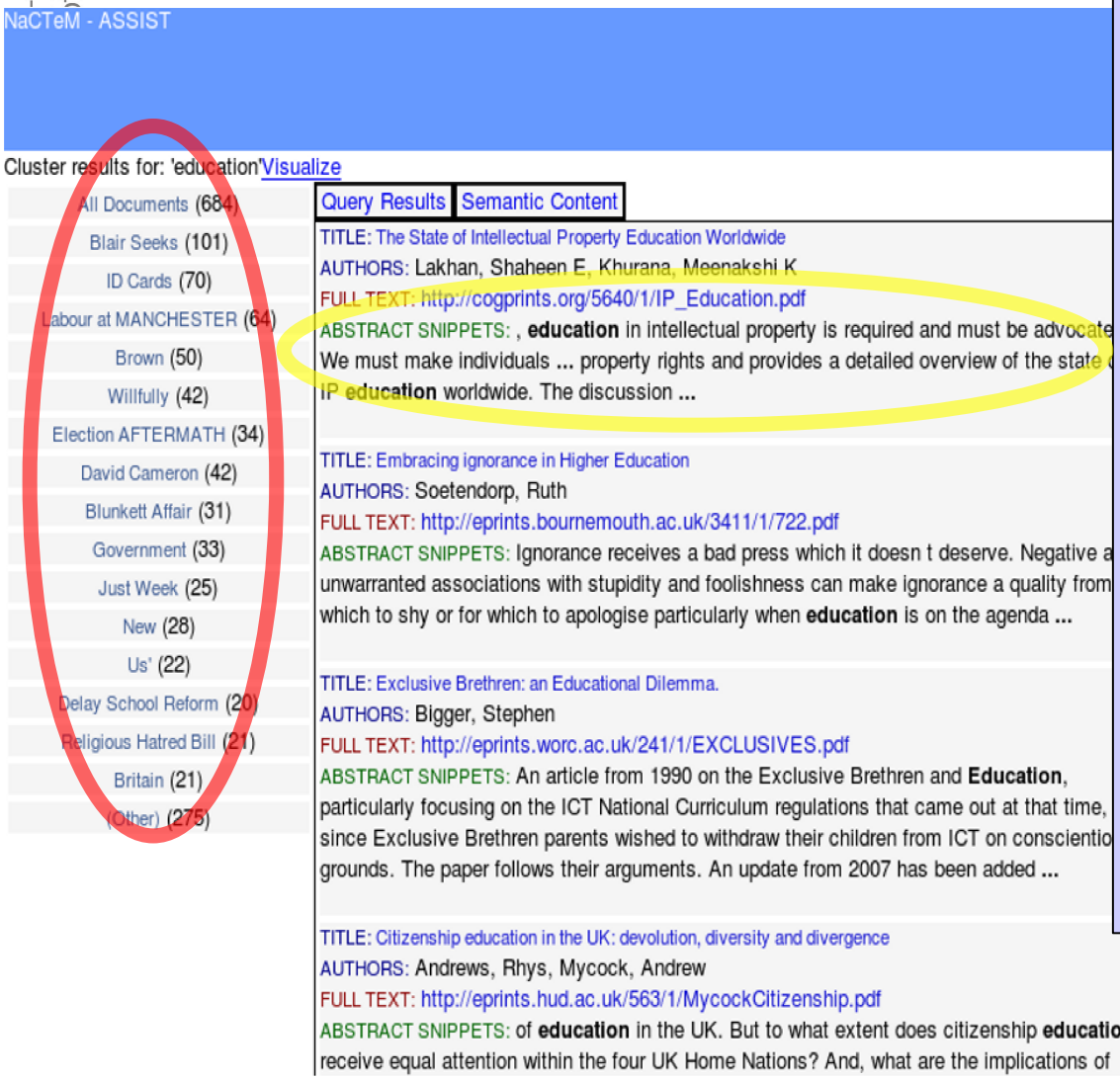
✓ Browsing documents through a domain taxonomy

Improving the rank of query results

• Resolution of Pronominal Anaphora relations to compute the real frequency of search words

(*e.g.* The dog eats the cat. It sleeps now)

Search Result Interface



- ✓ Clustering the query results in real time
- Lingo algorithm merges instances of commonly occurring phrases, keeping the best candidate to describe each cluster
- ✓ A familiar presentation of query results including snippets

Search Result Interface

Document content is described using semantic information

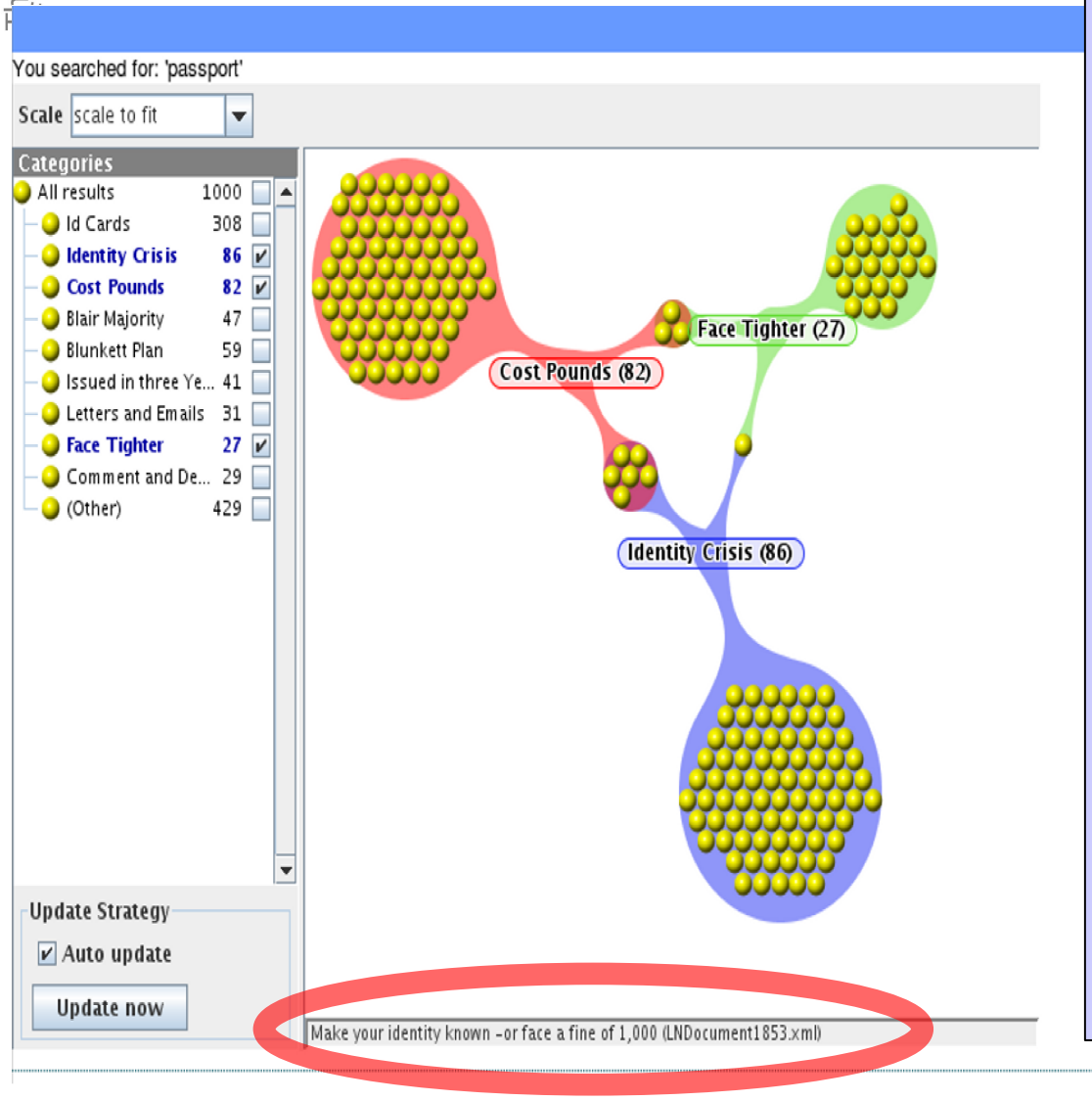
✔ makes document analysis easier, faster and more efficient

Access to document contents

Document content is described using semantic information

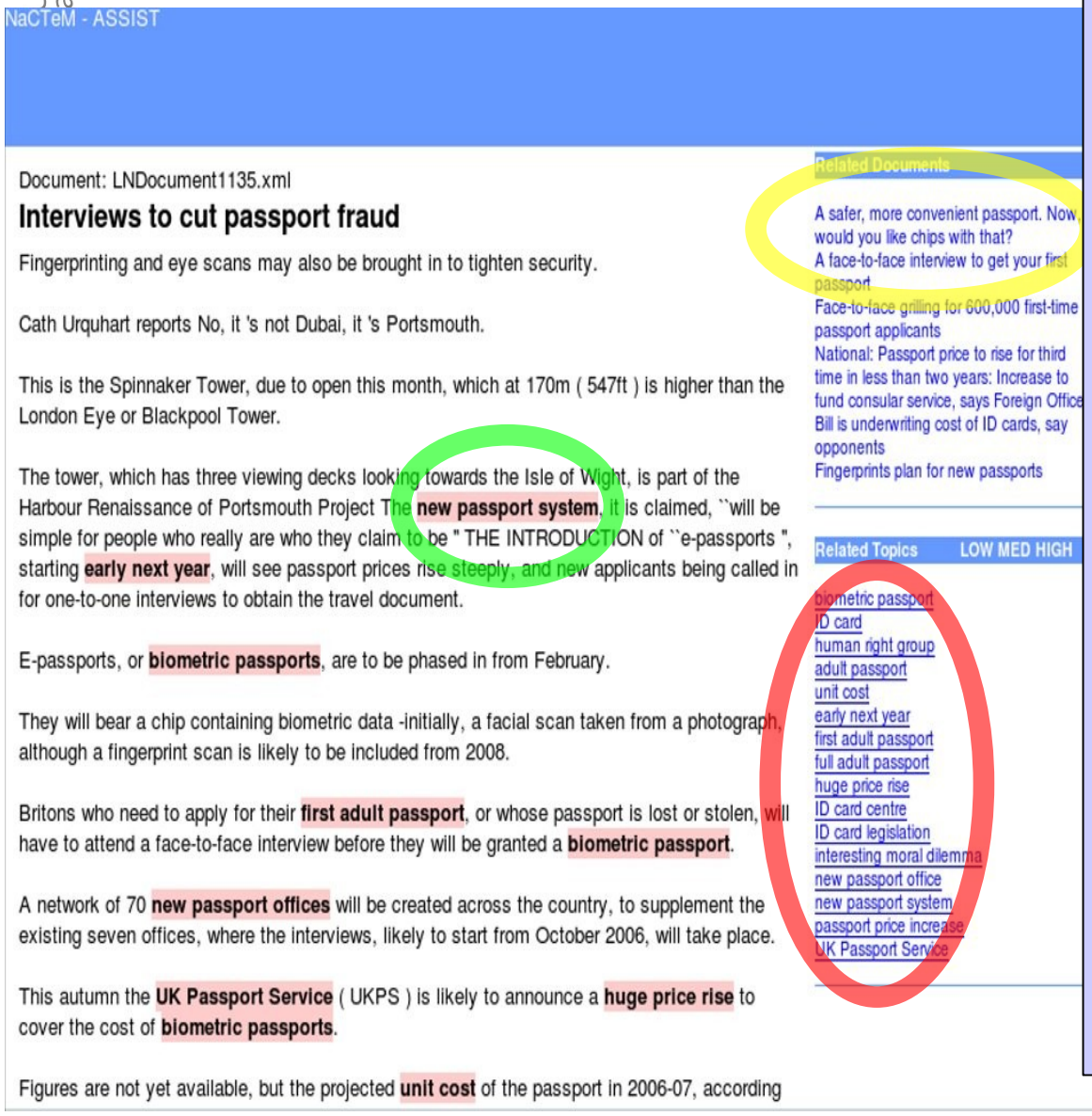
- Metadata: informing the origin of documents
- Terms: most significant multi-words phrases in the document
- Named Entities: main discourse objects belonging to predefined categories
- Lexical chains: gathering terms to build up concept representations

Query Results Visualization



- Examination of cluster memberships via a friendly visualisation interface
- Graphical representation of the intersection between the clusters provides immediate visualization of cluster relations
- ✔ Information regarding membership of particular cluster

Document Analysis



- ✓ Identification of conceptually similar documents using the most commonly occurring terms and words in the source document
- ✓ Highlighting selected semantic information within the document
- ✔ Selecting terms according to their importance and using them to browse documents

MANCHESTER 1824

The University of Manchester

Conclusion

NaCTeM
The National Centre for Text Mining

- Both applications designed for annotating documents but TM software complements the CAQDAS software
- TM techniques help the fastidious annotation stage of the qualitative analysis
- Presentation of the ASSIST project for evaluating the benefits of a tool based on TM for frame analysis of Media