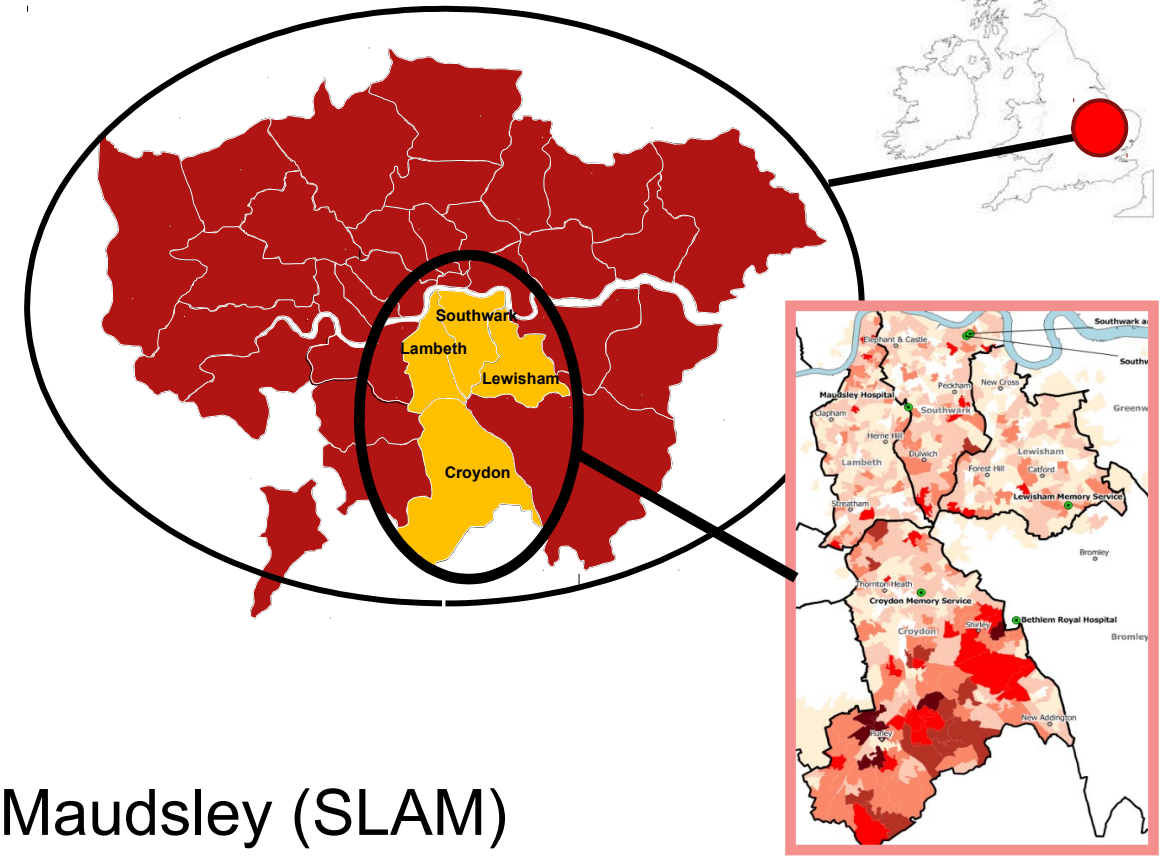
Bio-YODIE, Mimir and Prospector

Overview

- Bio-YODIE for semantic annotation and named entity linking
- Mimir for search
- Prospector for visualization

SLAM Biomedical Research Centre for Mental Health (BRC-MH)

King's College London (KCL)

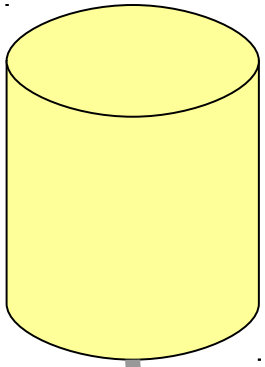


South London and Maudsley (SLAM)

• Context: EHR and search at SLAM

The Patient

- **Journey**
- **System (PJS)**



Coverage: Four London boroughs

Local population: c. 1.1 million

Clinical area: specialist mental health

Active patients: c. 35 000

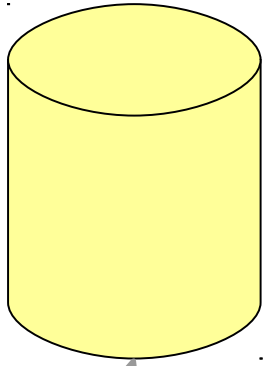
Total inpatients: c. 1 000

Total records: c. 250 000 people

Documents: 15 million (D-CRIS: 75 m)

CRIS

- **Interactive search**
- **FAST index**
- **SQL RDBMS**



CRIS and Information Extraction

... finishing A levels ..

.. made no eye contact ..

... MMSE=27 on 15 Feb ...

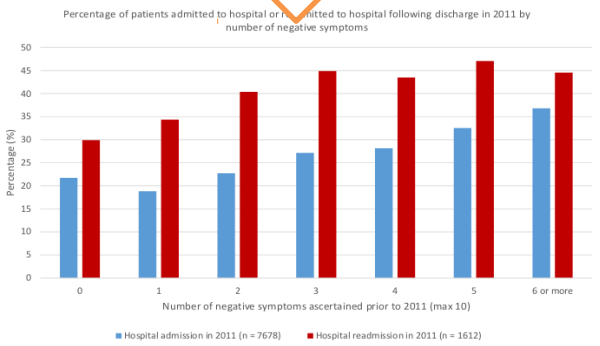
... meals on wheels ..

... smokes 20 a day ..

Extract

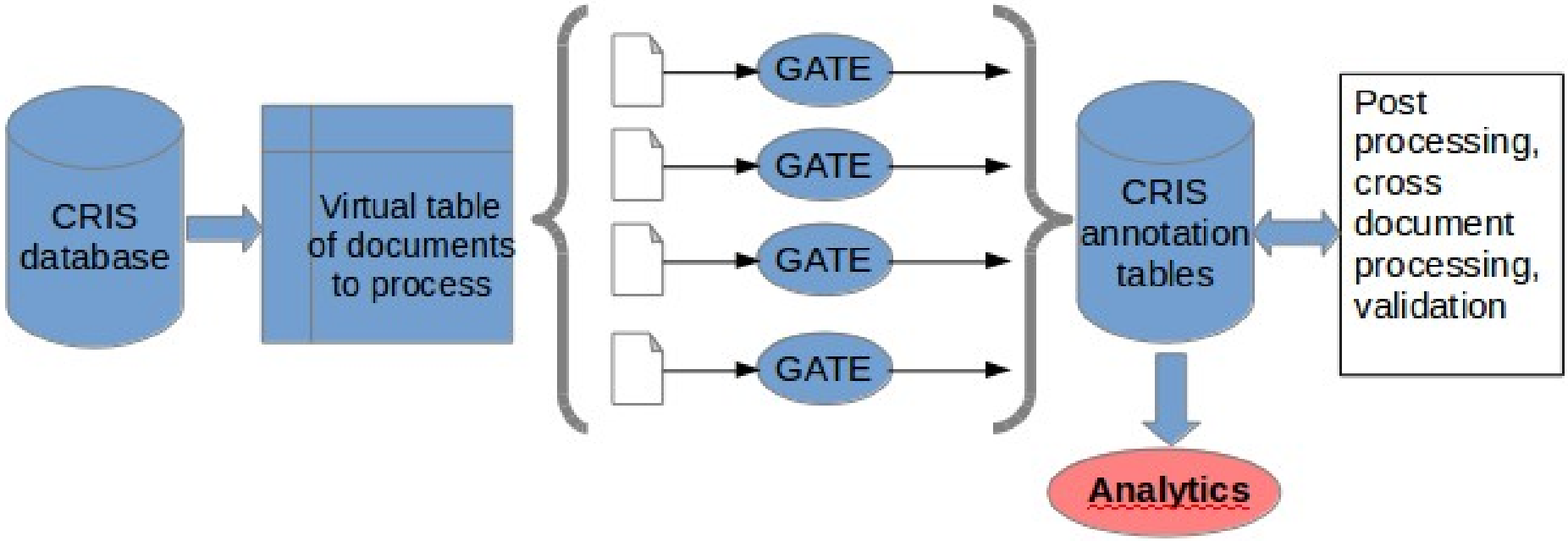
ID	Type	Date	Value
245	MMSE	15/02/16	27/30
8467	Education	12/03/16	2dary
2478	PANSS	15/04/16	
943	Smoking	10/01/16	Current
875	Social care	15/02/16	Current

Analyse



GATE @ SLAM

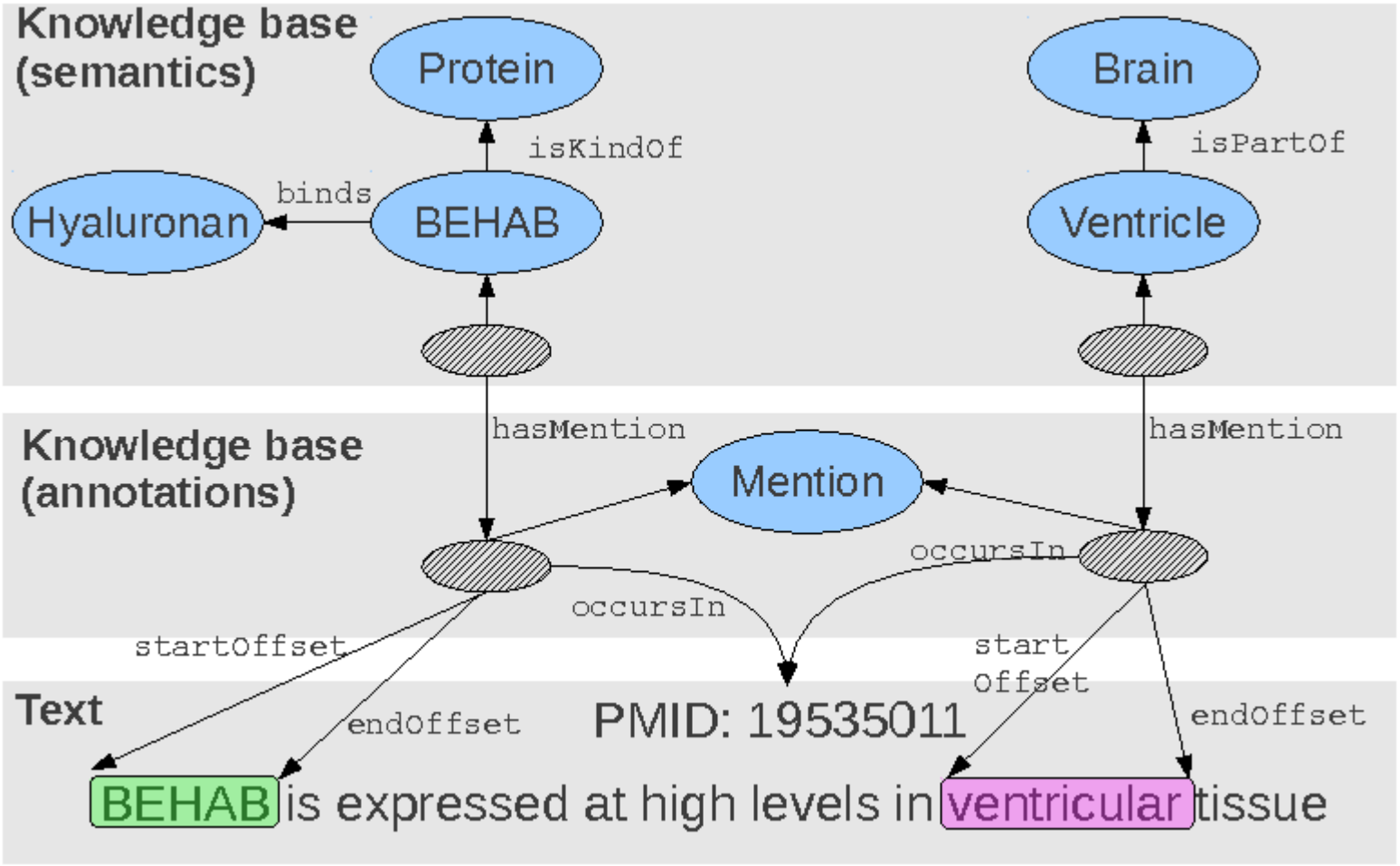
Over 60 GATE applications are routinely run over 15 million patient documents, to extract structured information



GATE @ SLAM

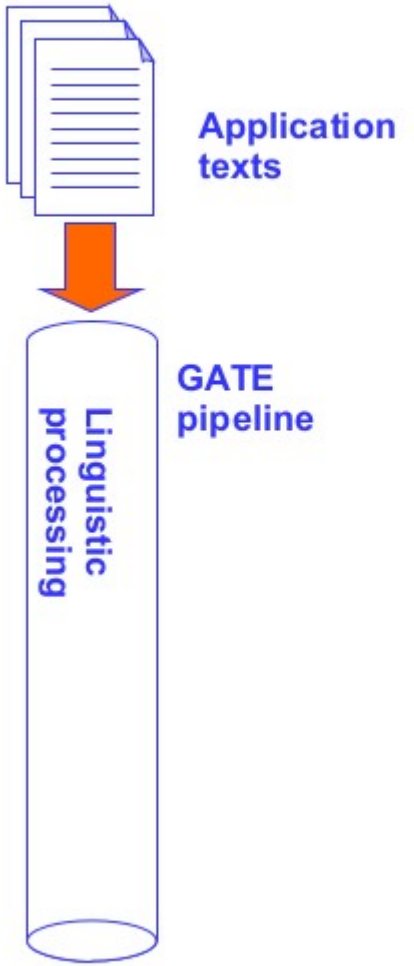
Application group	Application sub-group
Symptoms	positive, negative, disorganisation, manic, catatonic, affective, behaviour
Background	cognitive function, social care, living alone, education, smoking, HIV status, physical disorders, illicit substances
Intervention	pharmacotherapy, psychotherapy
Outcomes	trajectories, adverse events

Why semantic annotation?

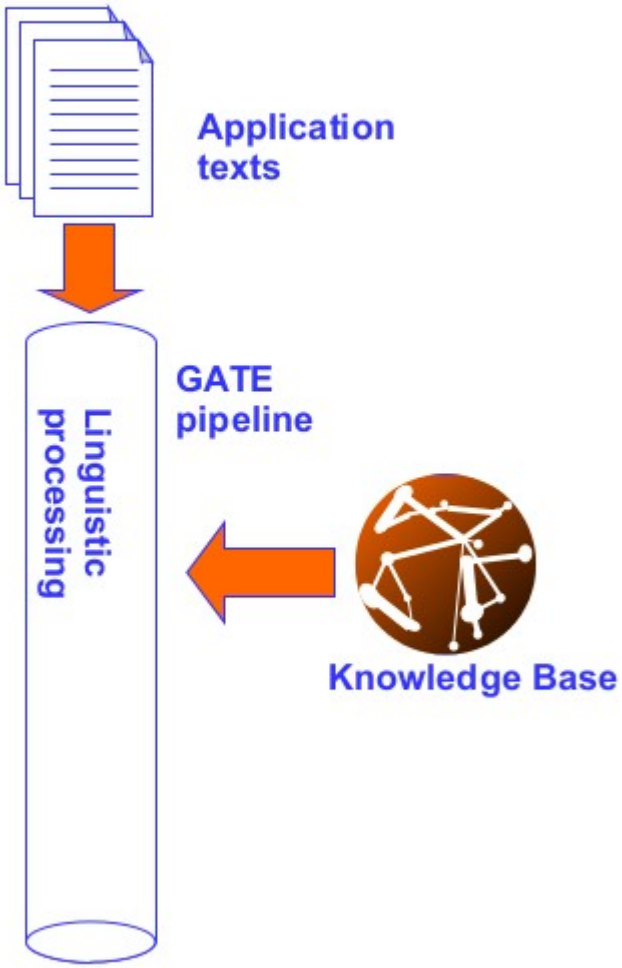


A deeper understanding of text ..

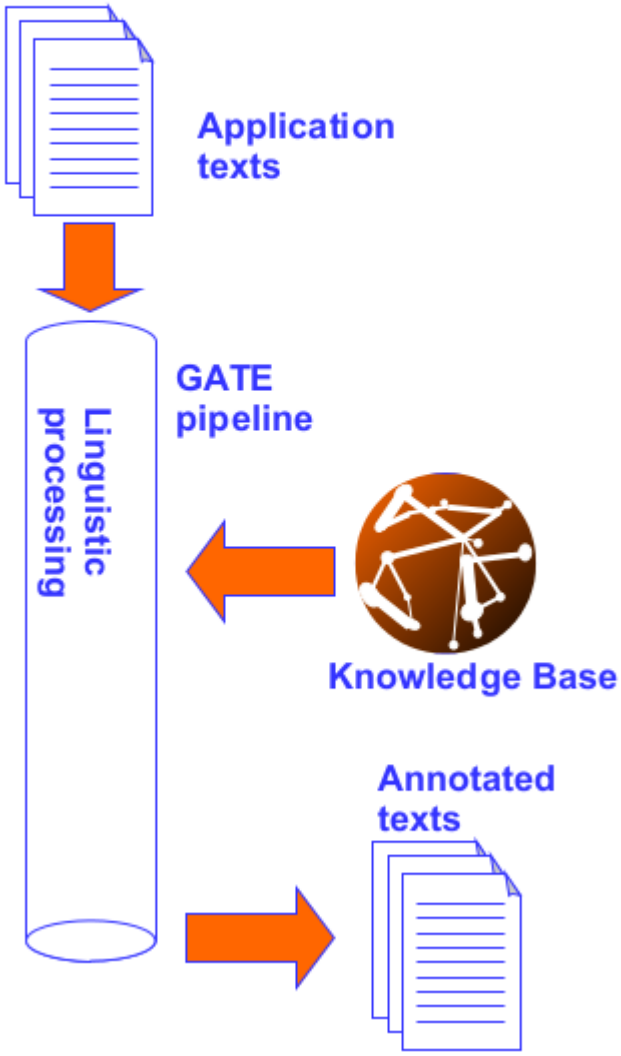
Linking text and knowledge



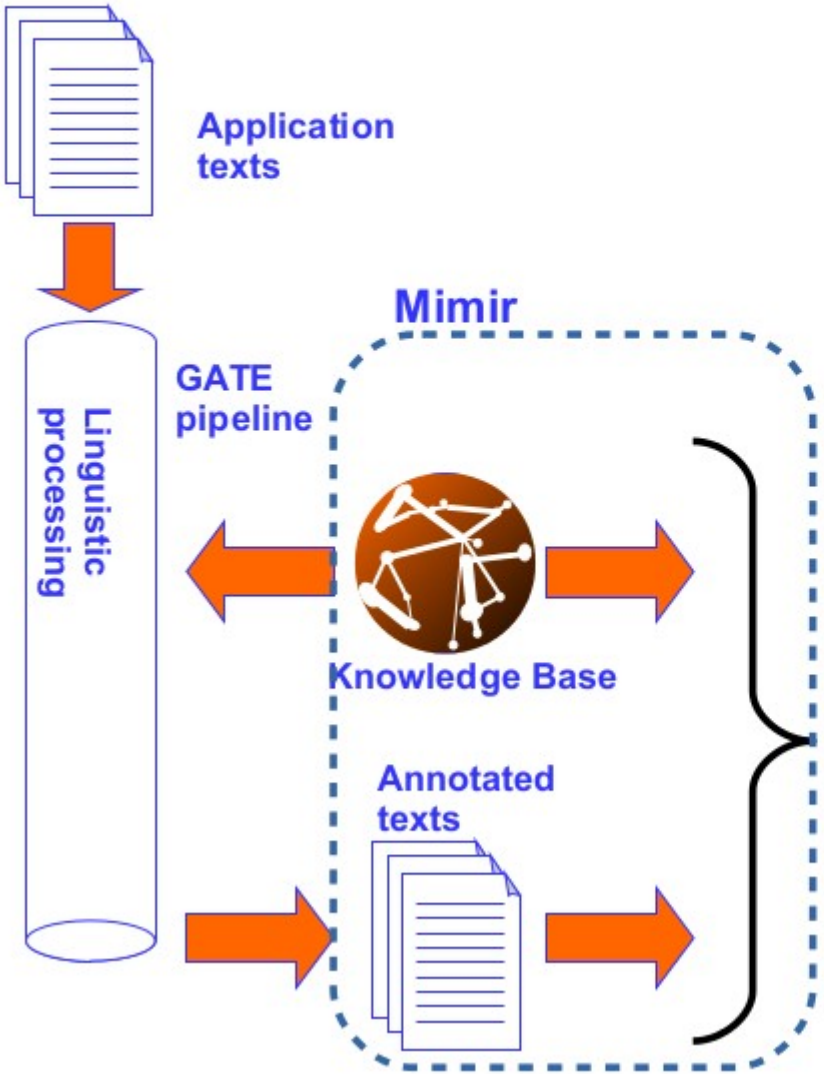
Linking text and knowledge



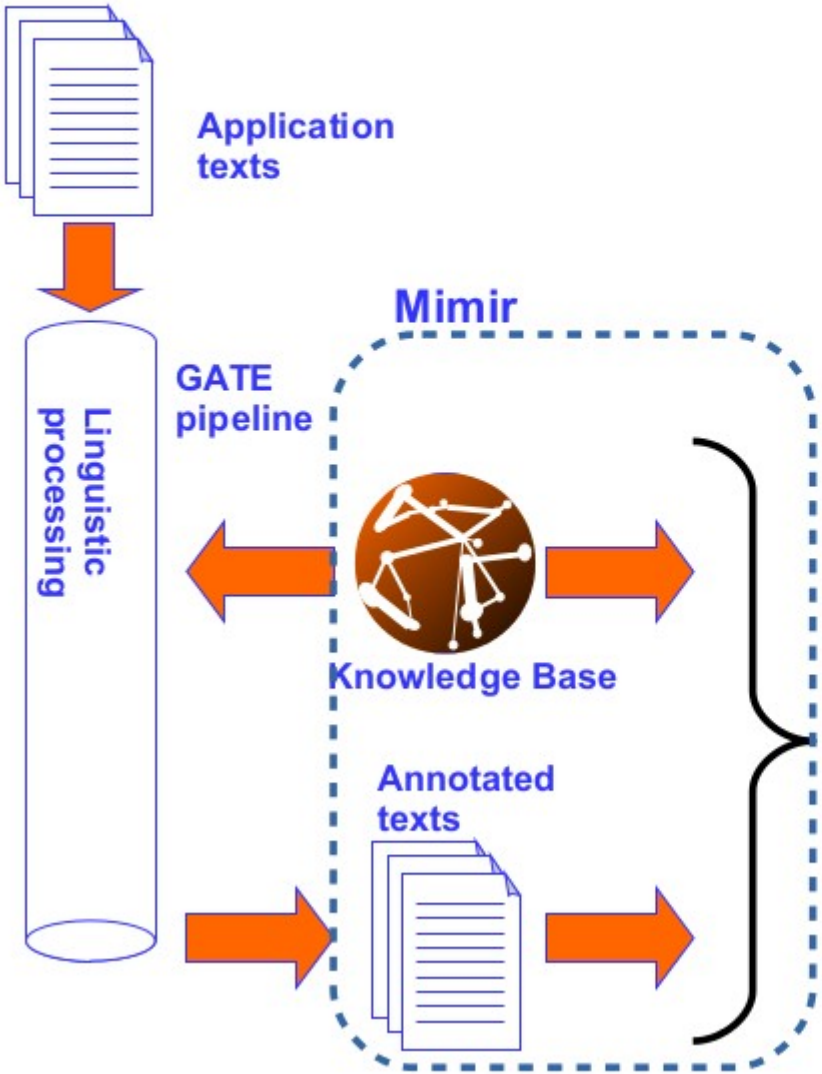
Linking text and knowledge

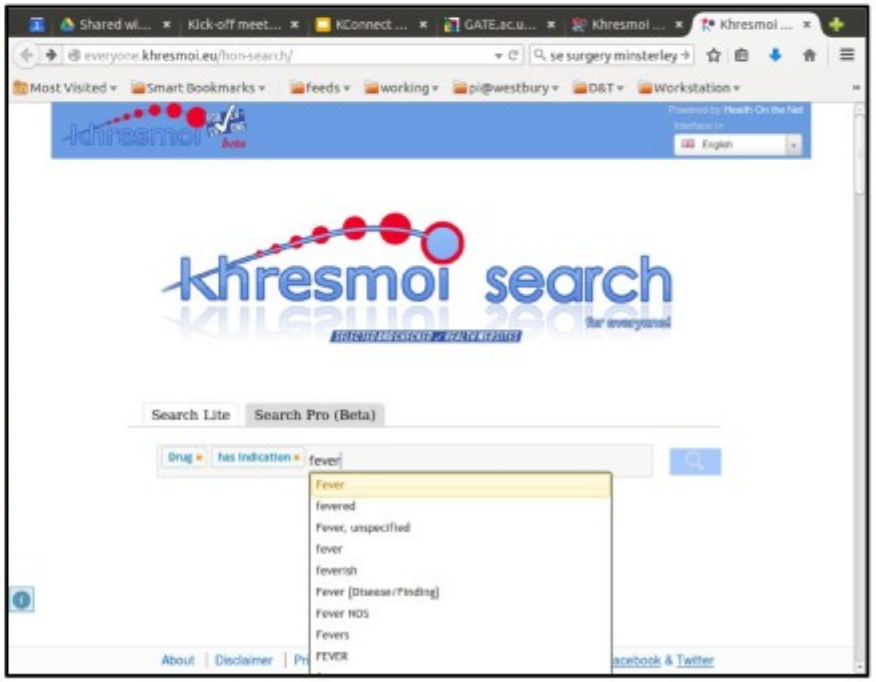


Linking text and knowledge

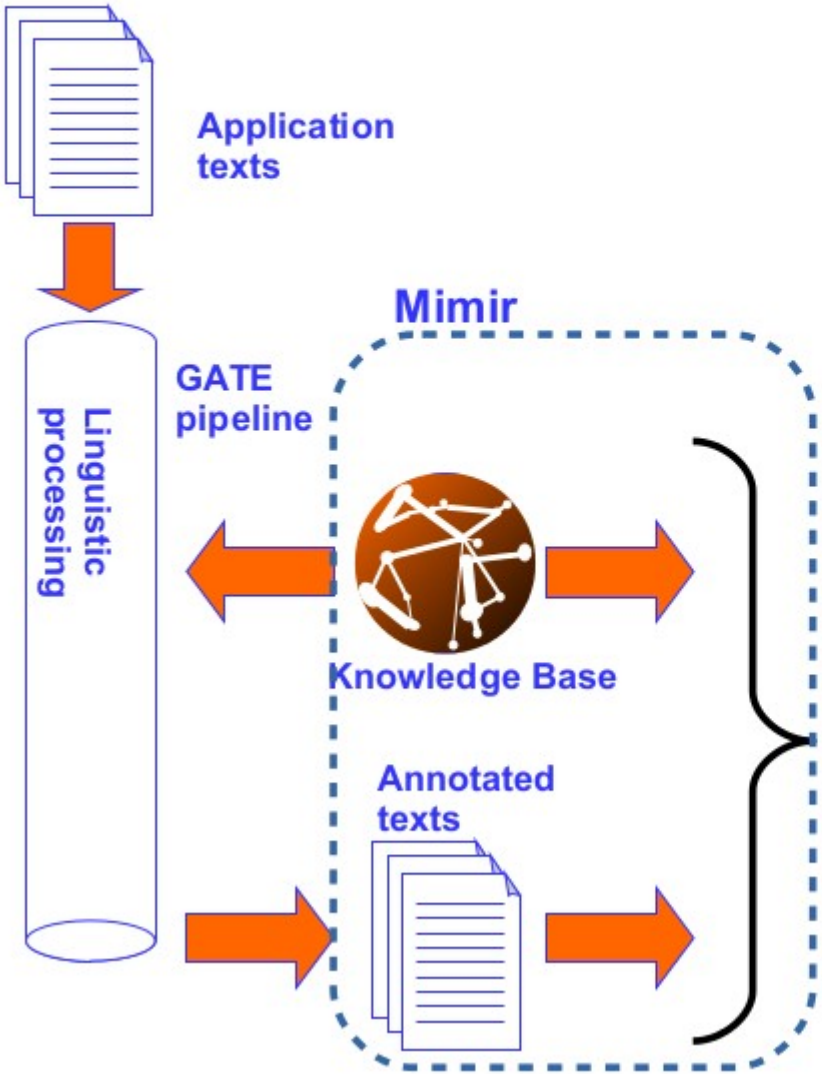


Linking text and knowledge

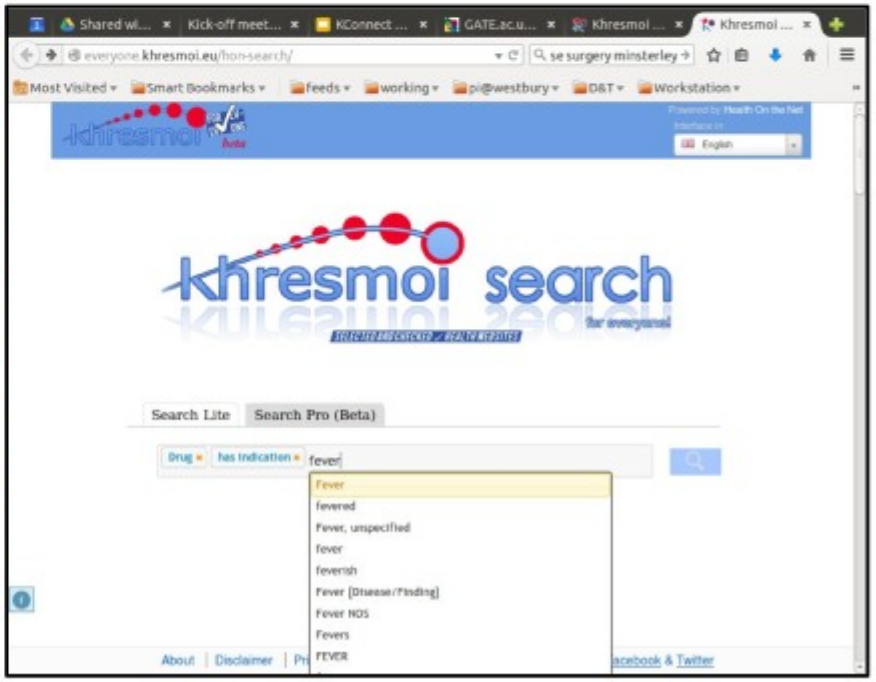




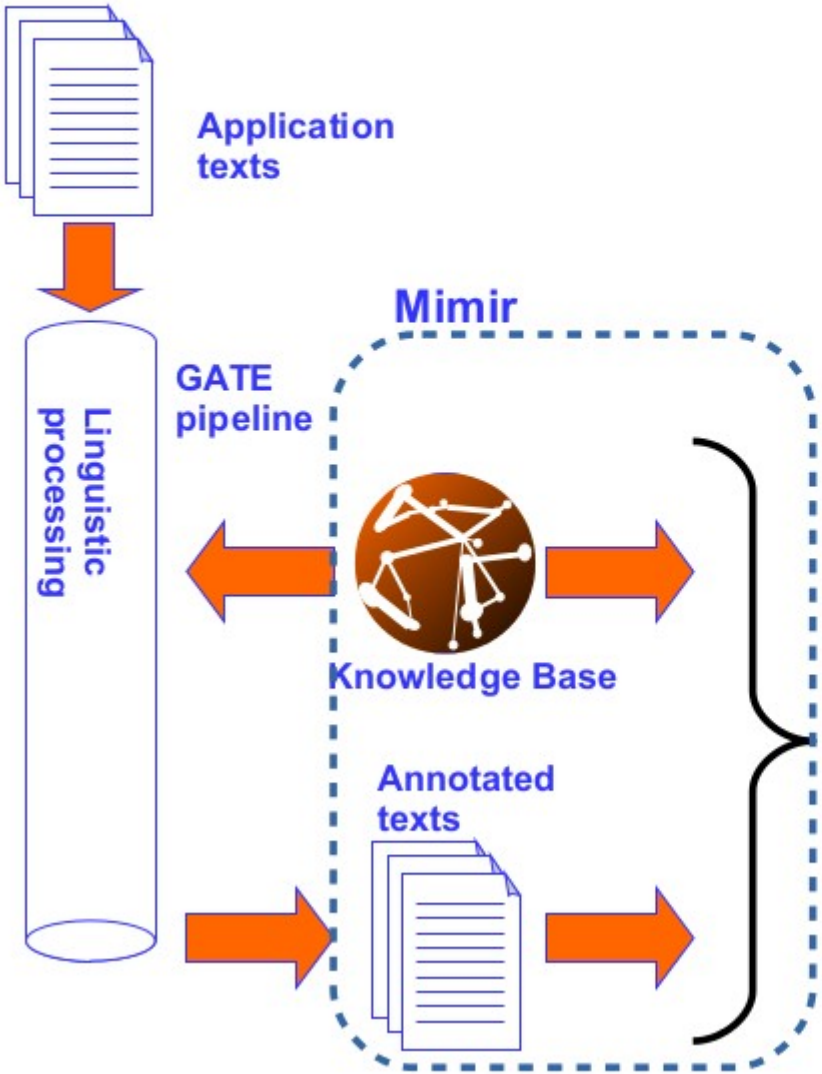
Linking text and knowledge

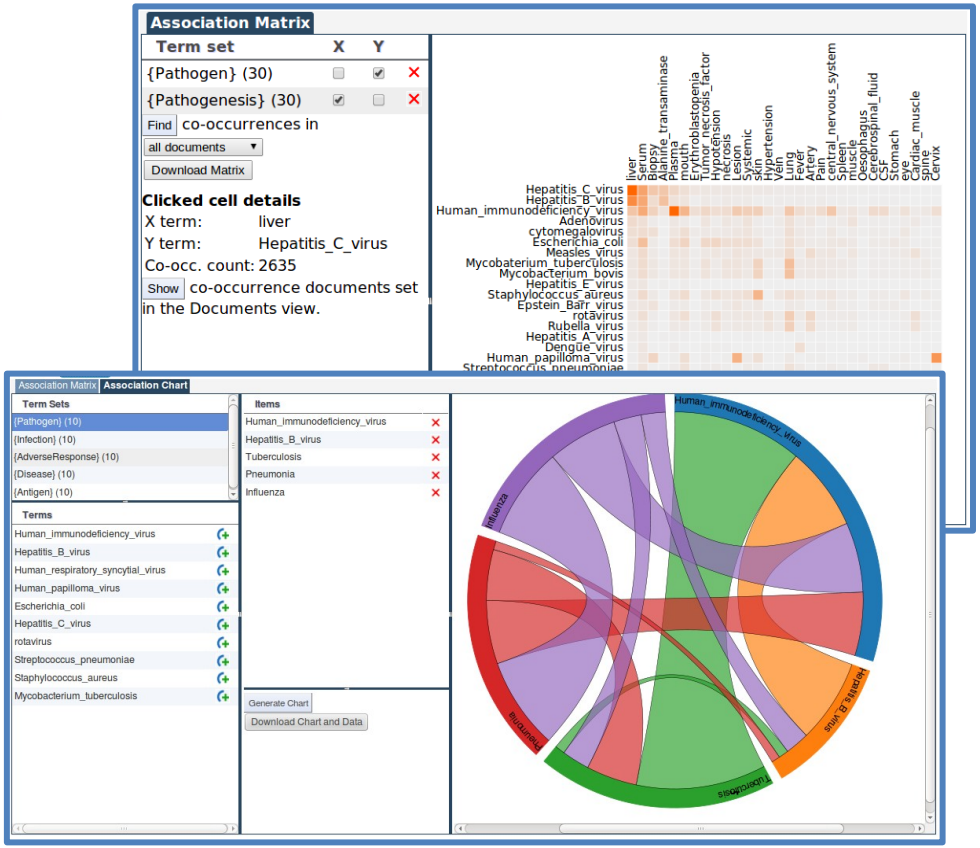


Find patient records that mention drugs used to treat disorders commonly associated with depression

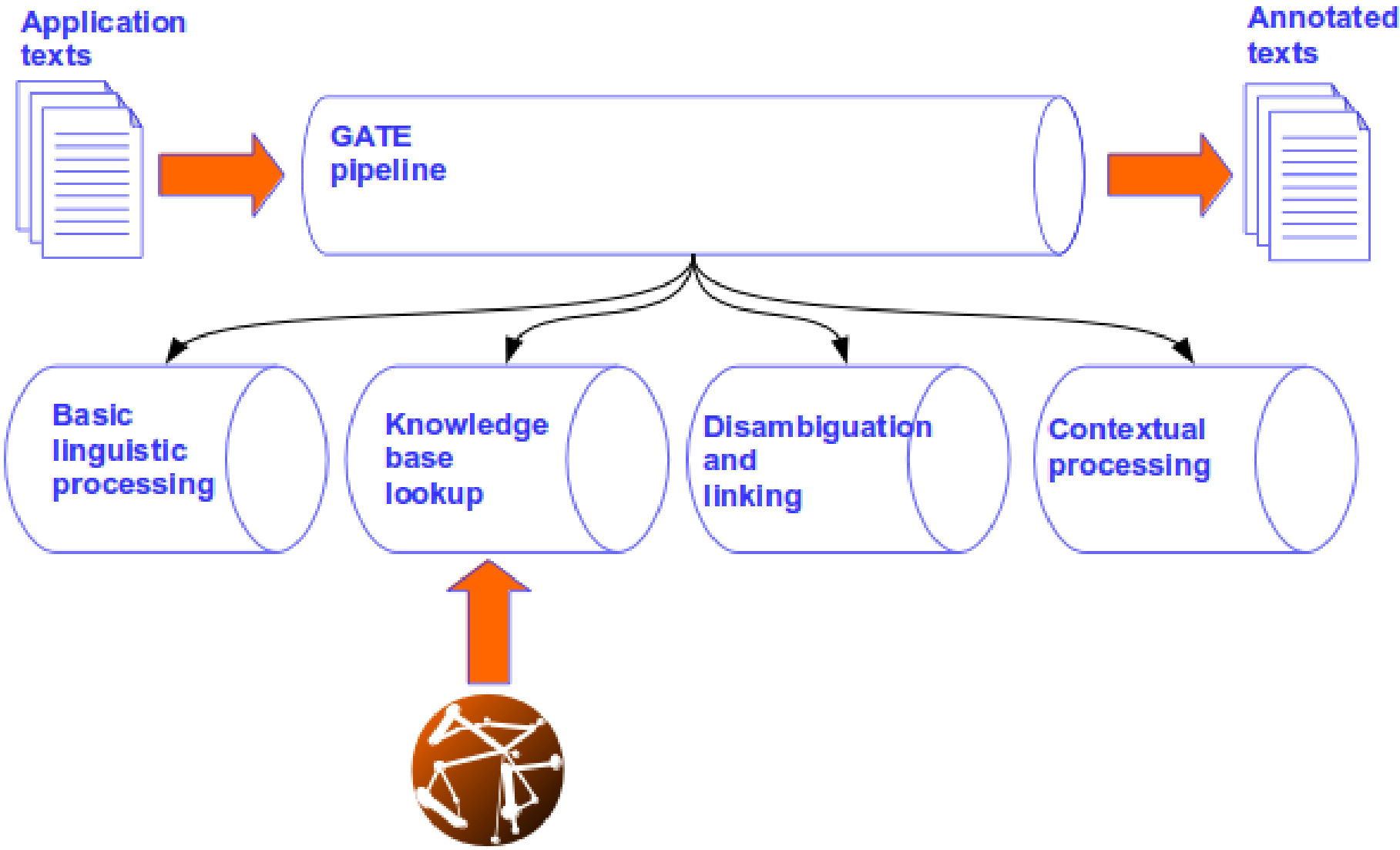


Linking text and knowledge





Bio-YODIE


Application texts
GATE pipeline
Annotated texts
Basic linguistic processing
Knowledge base lookup
Disambiguation and linking
Contextual processing

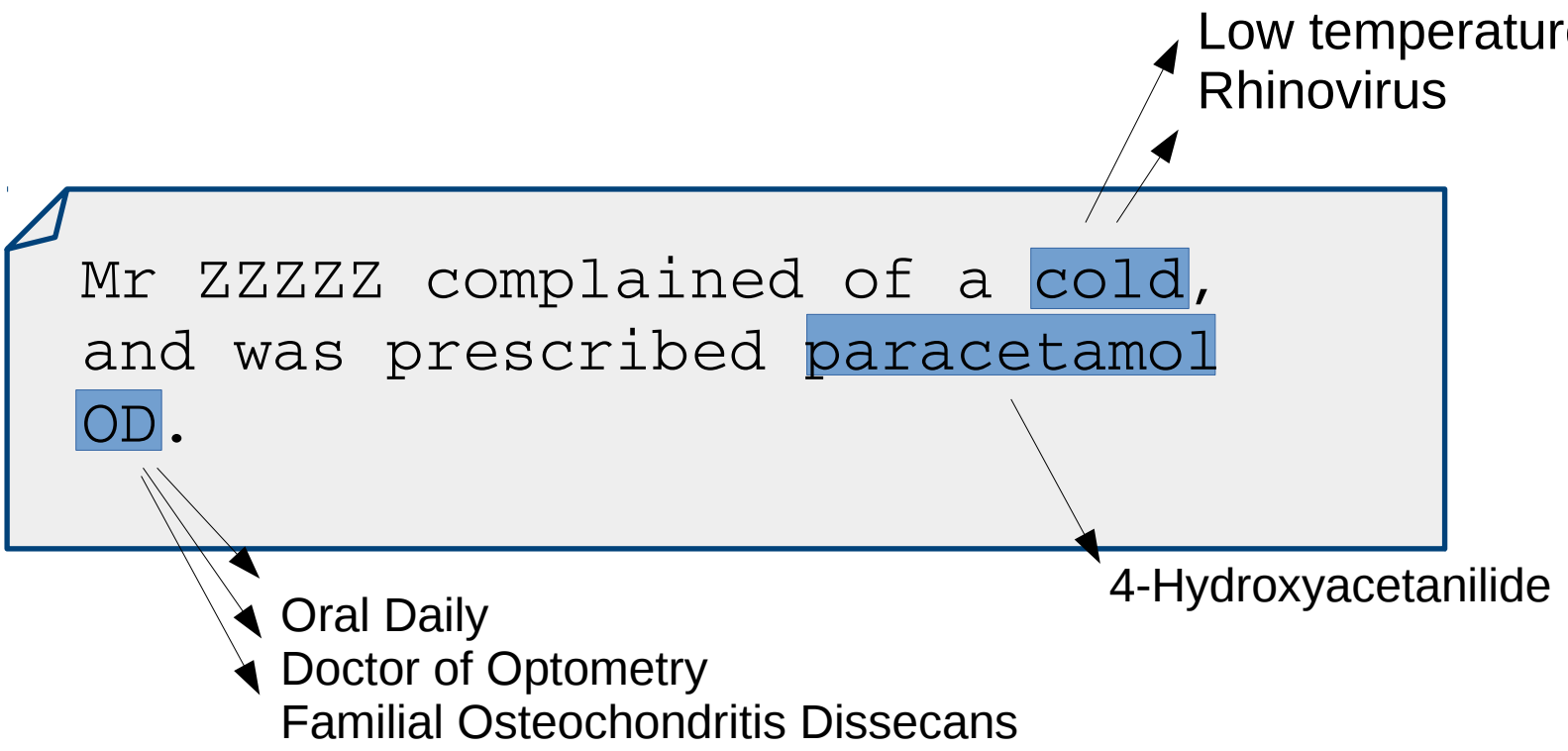
Named Entity Linking

- A central task in semantic annotation
- Connect mentions in text to their referent in a knowledge base, so we can find out about them and coreference them
- AKA term identification and mapping, or normalization

Mr ZZZZZ complained of a cold, and was prescribed paracetamol OD.

Named Entity Linking

- You have found mentions in text, but there are several possibilities for what something is referring to
- This is the disambiguation problem



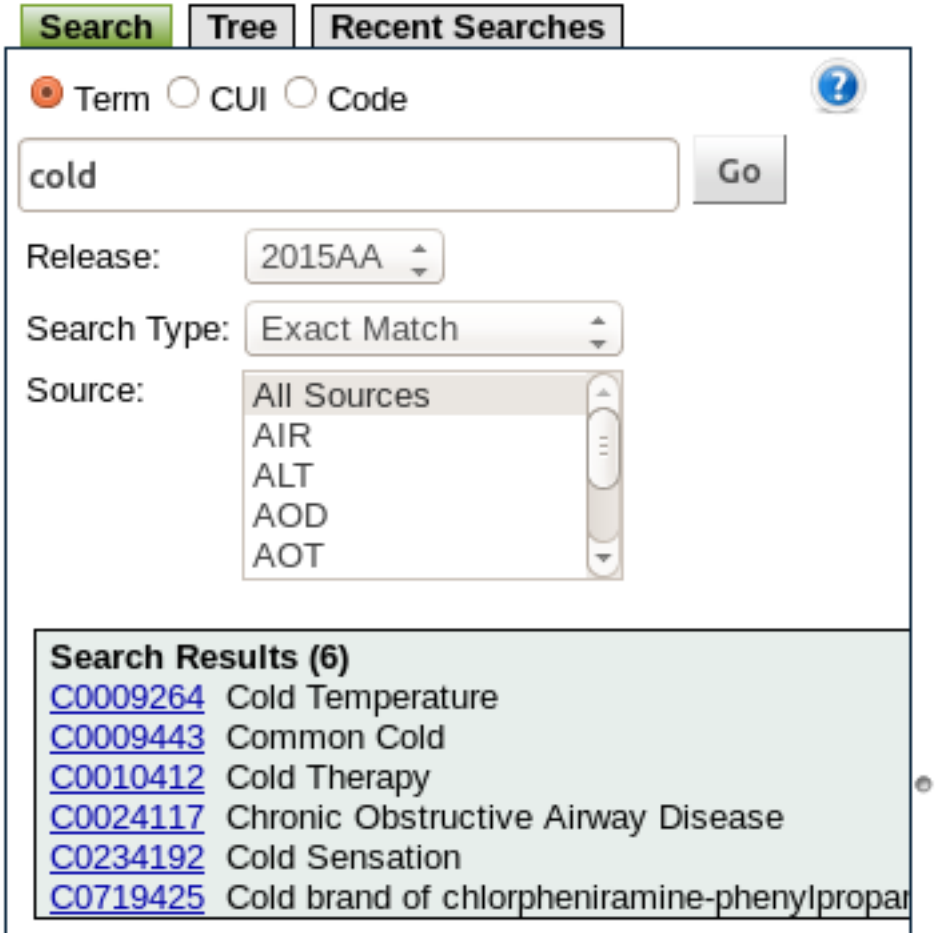
Bio-YODIE

- Dictionary based term identification based on names of UMLS entities of relevant types
- Retrieval of possible interpretations—any entity that can be called that
- Multiple disambiguation strategies gauging how well the candidates fit
- Scores weighted based on a supervised model to choose the best

UMLS Metathesaurus

- Unified Medical Language System (from the US National Library of Medicine)
- Contains over two million biomedical and health-related concepts
- Combines many thesauri (“source vocabularies”)
- Much ambiguity! E.g. “unknown” is associated with 39 concepts
- The NLP “view” (subset) is more pragmatic, excluding unhelpfully ambiguous and spurious concepts

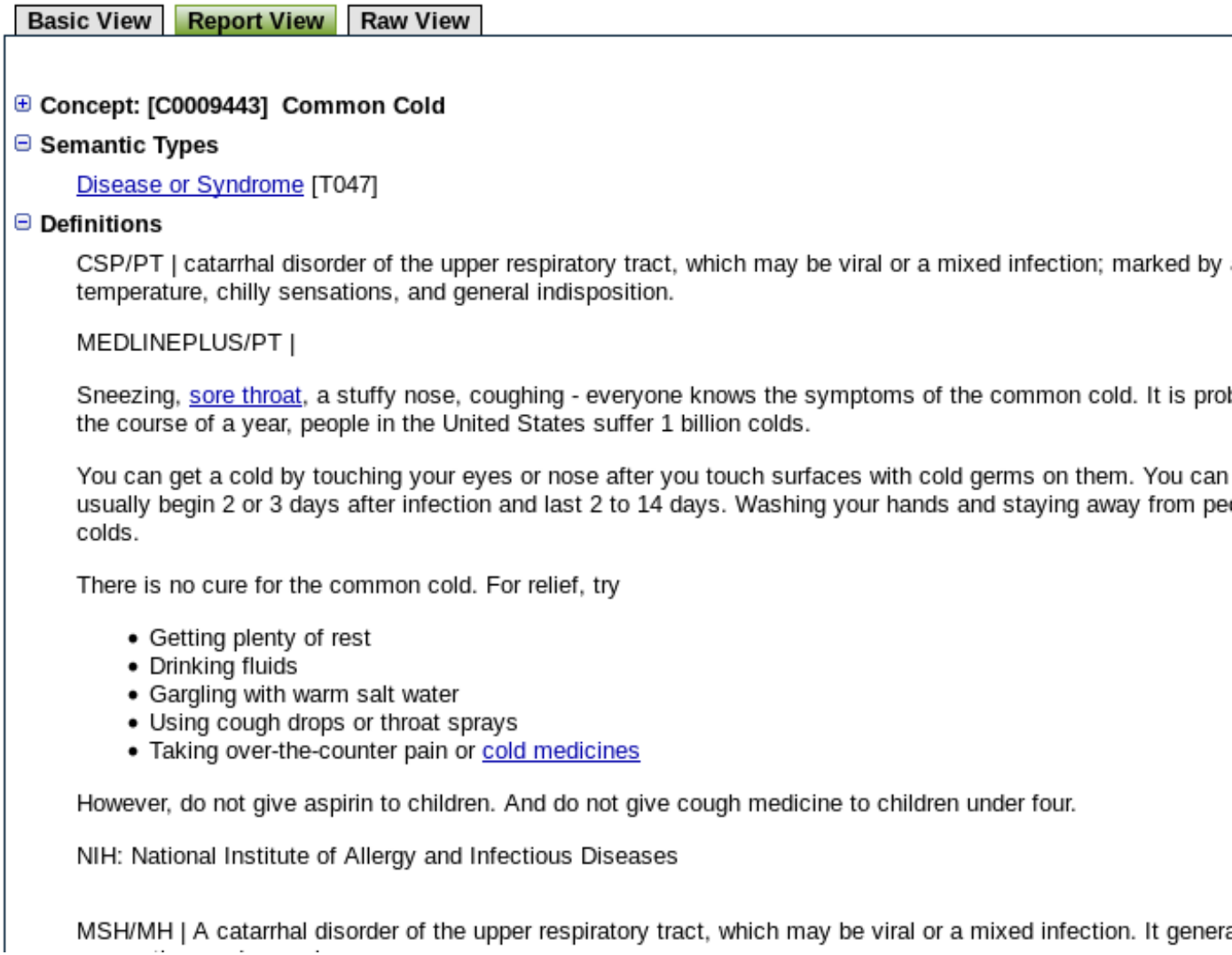
UMLS Metathesaurus



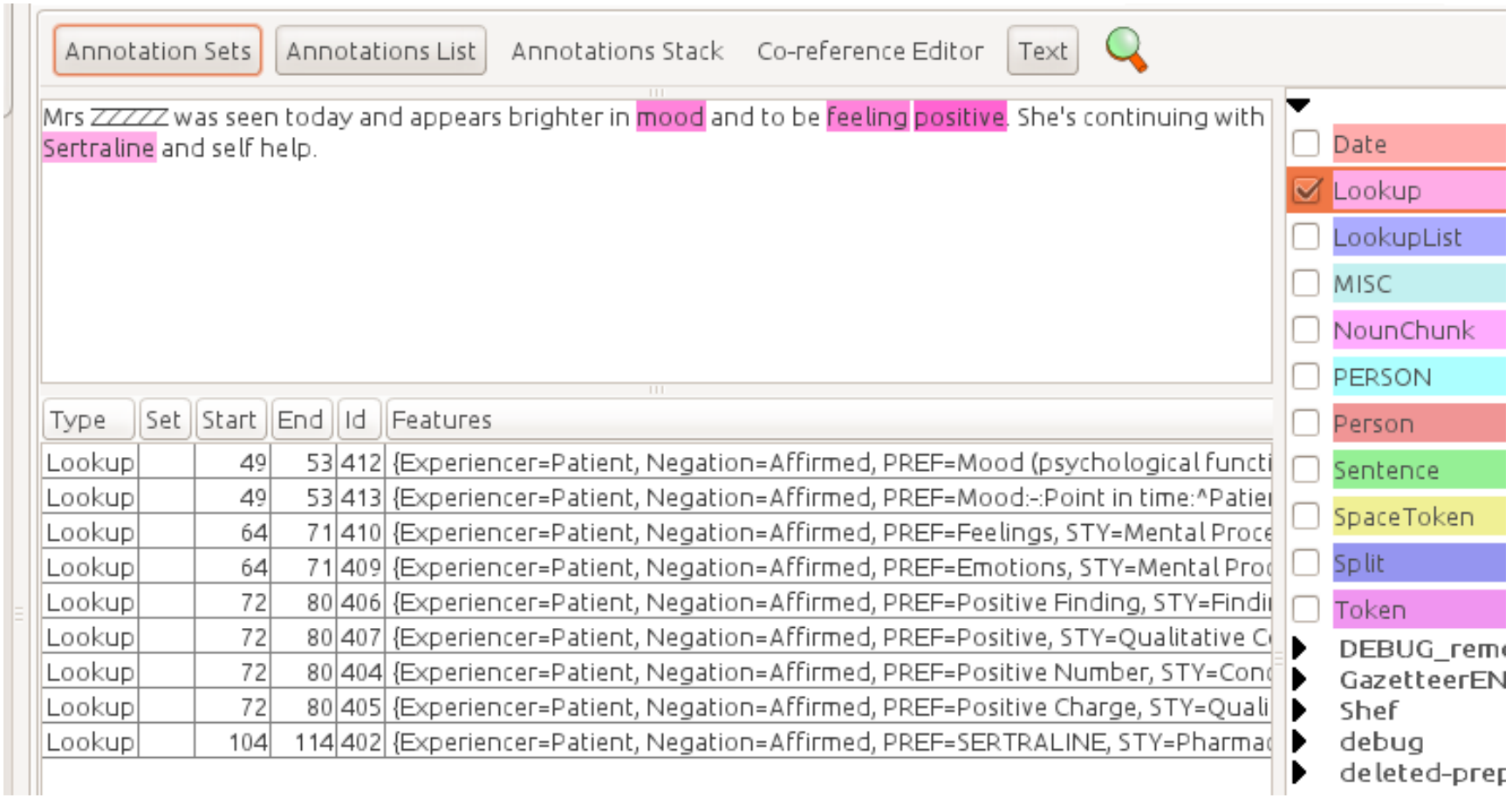
- “Cold” has six associated concepts
- “CUI” means concept unique identifier and is the unique code for the concept

UMLS Metathesaurus

- A concept has a type (in this case, “Disease or Syndrome”)
- It has definitions from different vocabularies
- It also has concept relations, i.e. things it is related to. “Common Cold” is related to “Respiration Disorders” for example



Bio-YODIE: Finding Mentions in Text



Bio-YODIE disambiguation scoring strategies

Strategy	Implementation	Example
Prior likelihood of entity mentions	MeSH term frequency in psychiatry journals	For "OD", the [overdose] entity is more common than [osteochondritis dissecans]
Ranked likelihood of co-occurrence	PageRank across a UMLS co-occurrence table	[overdose] is more likely to be mentioned with [depression] than is [osteochondritis dissecans]
Context similarity to entity definitions	UMLS definitions table	The context of "OD" is more likely to be close to the context of the [overdose] definition
Count of direct and indirect relations between entities	UMLS relations table	[overdose] is likely to have more links to [depression] than [osteochondritis dissecans] will have

Preliminary results

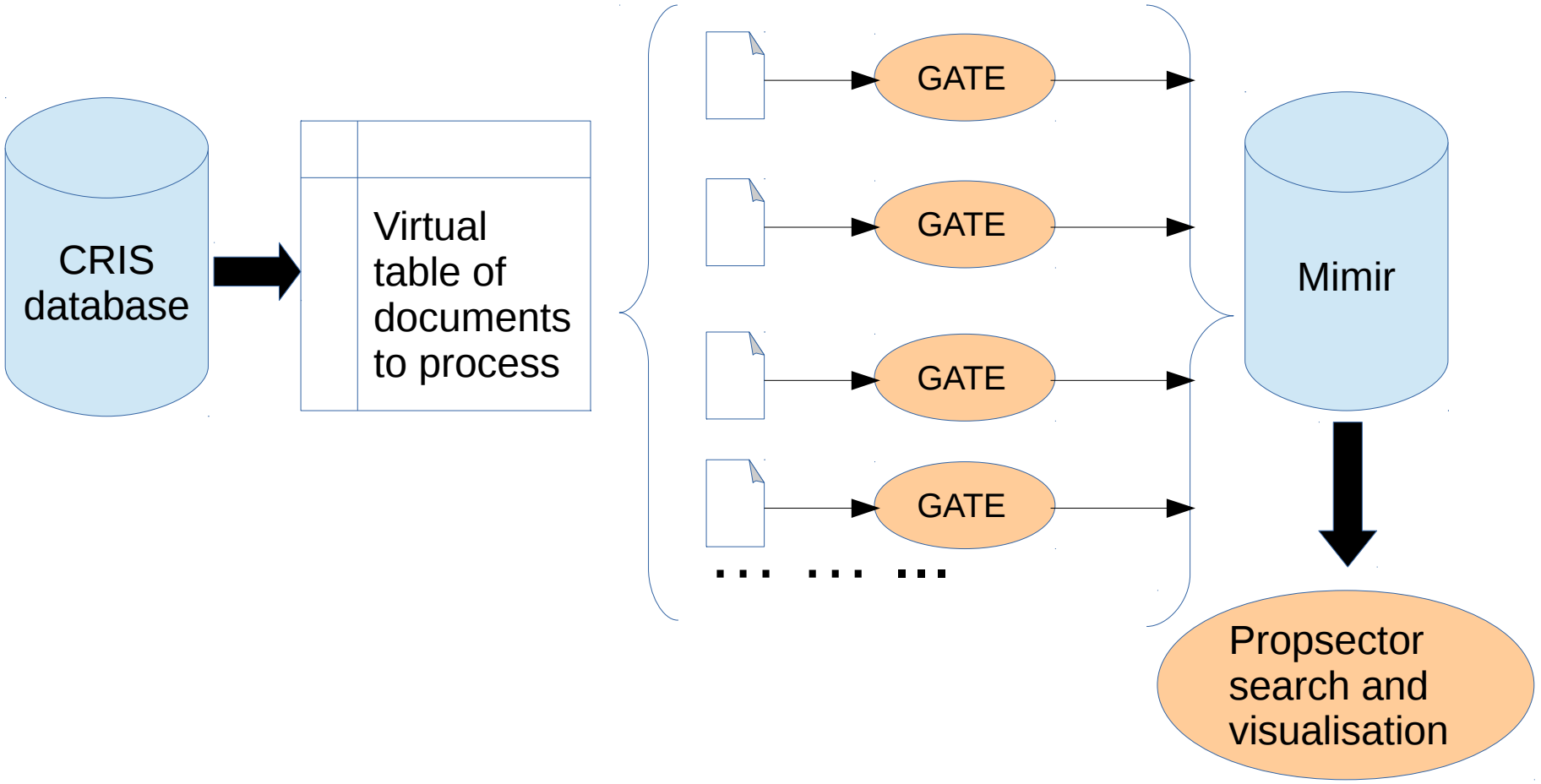
	Total correct	Correct rejections	Correct accepted candidates	Accuracy
Always accept one candidate at random	410	0	410	0.50
Accept one random candidate 63% of the time, reject all candidates otherwise	358	100	258	0.43
Bio-YODIE, reject if all scores are zero	488	82	406	0.59
Bio-YODIE, rejection threshold 0.05	435	142	292	0.53

Results obtained on 826 mention manually annotated CRIS corpus

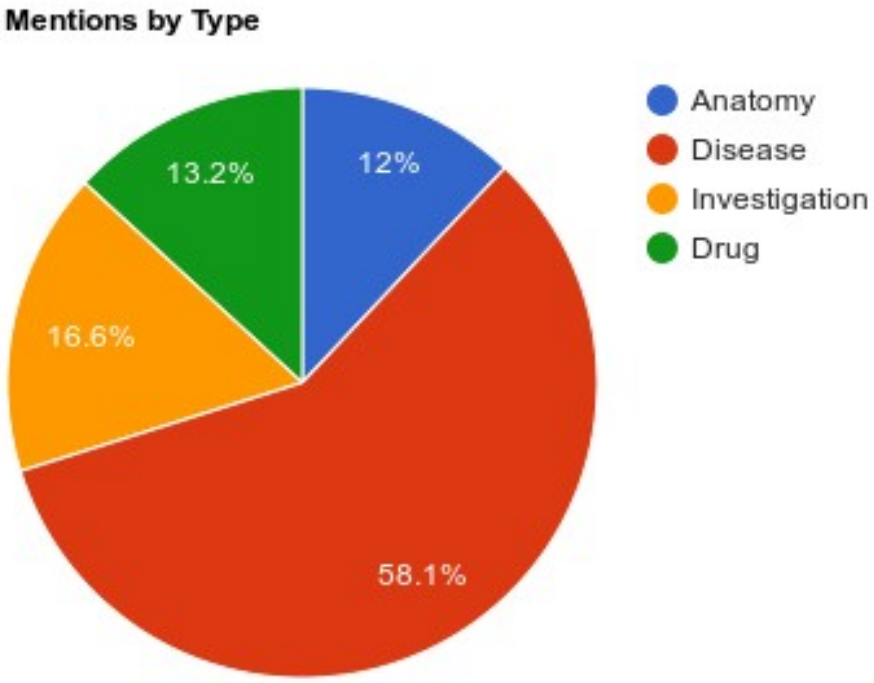
Beyond named entity linking

- Context—included
  - Based on the popular NegEx algorithm
  - Negation
  - Subject
  - Historicity
- Time – temporal expression extraction—still to be included
  - Supervised models (SVM)
  - Pattern matching grammars for imprecise times
- Language – now available for French

Deployment and indexing

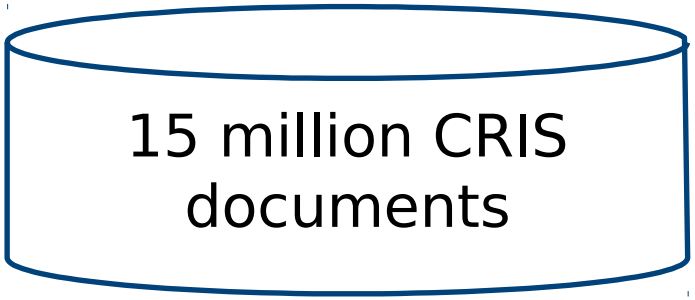


Mimir index statistics



Proportion of types found on 50K CRIS test documents (patient records)

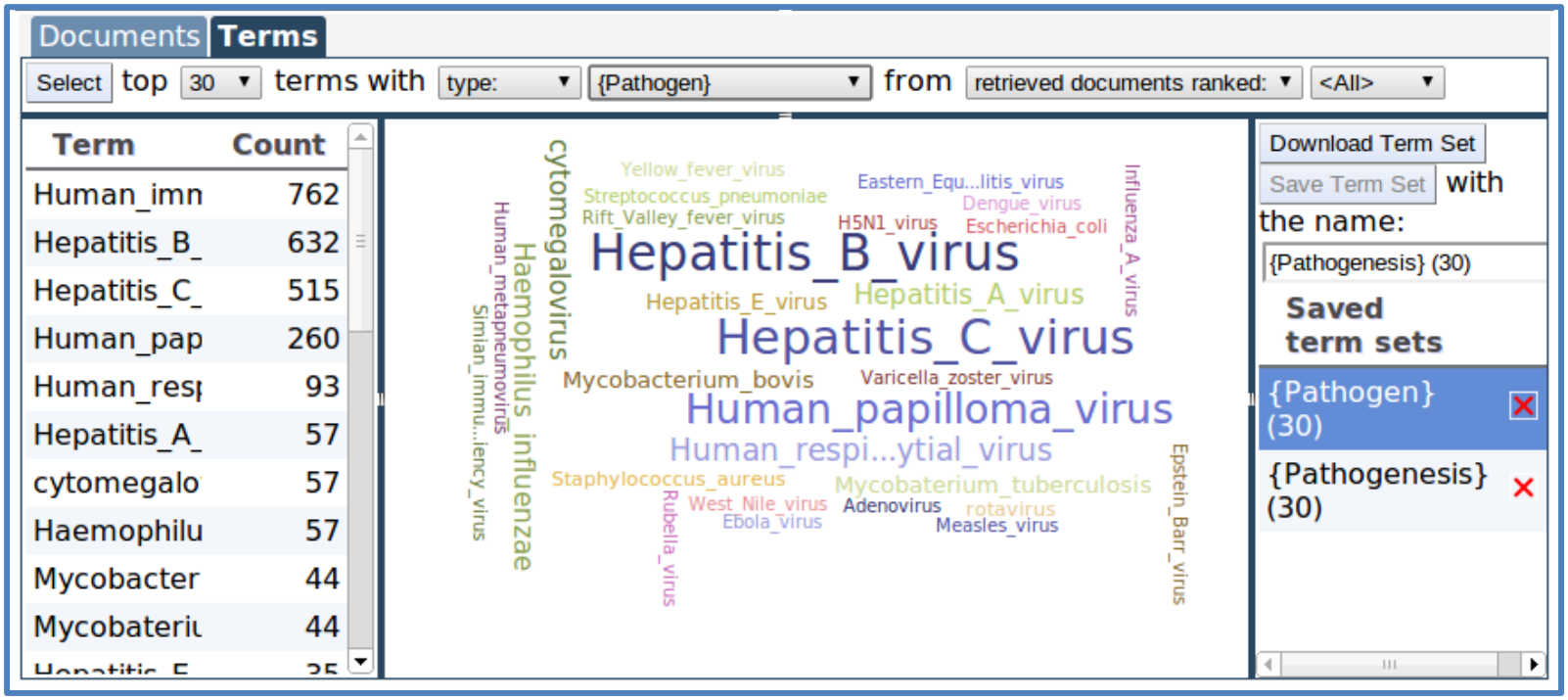
- Using GCP, we (USFD and KCL) have indexed ...



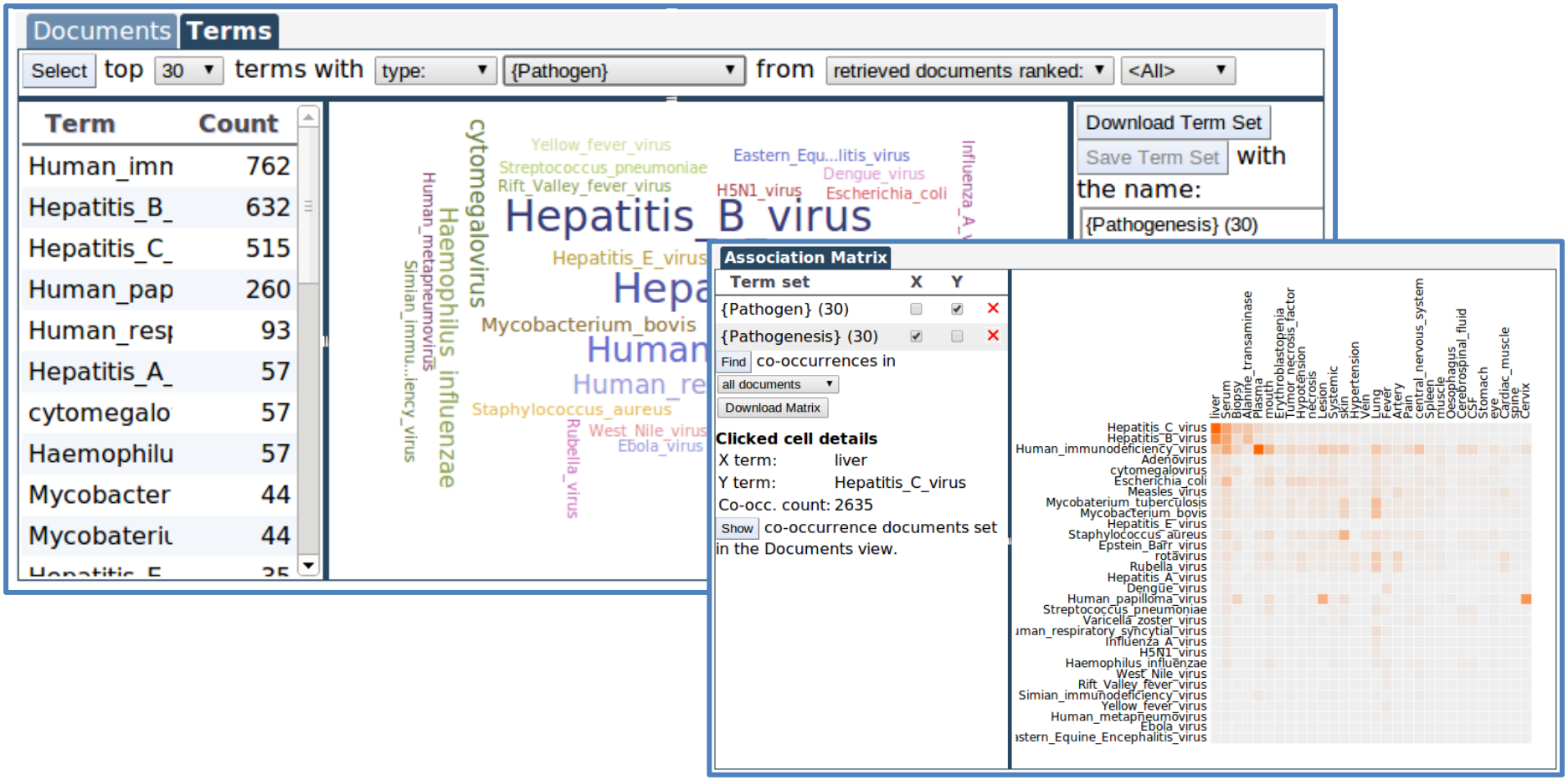
Prospector

- Mimir provides a semantic search interface allowing complex queries
  - E.g. find mentions of mood disorders occurring in the same patient record as heart conditions
- Prospector is Mimir's visualization layer
  - Visualize relationship between heart conditions and mood disorders
- Integrates D3 to allow visualization of complex semantic queries

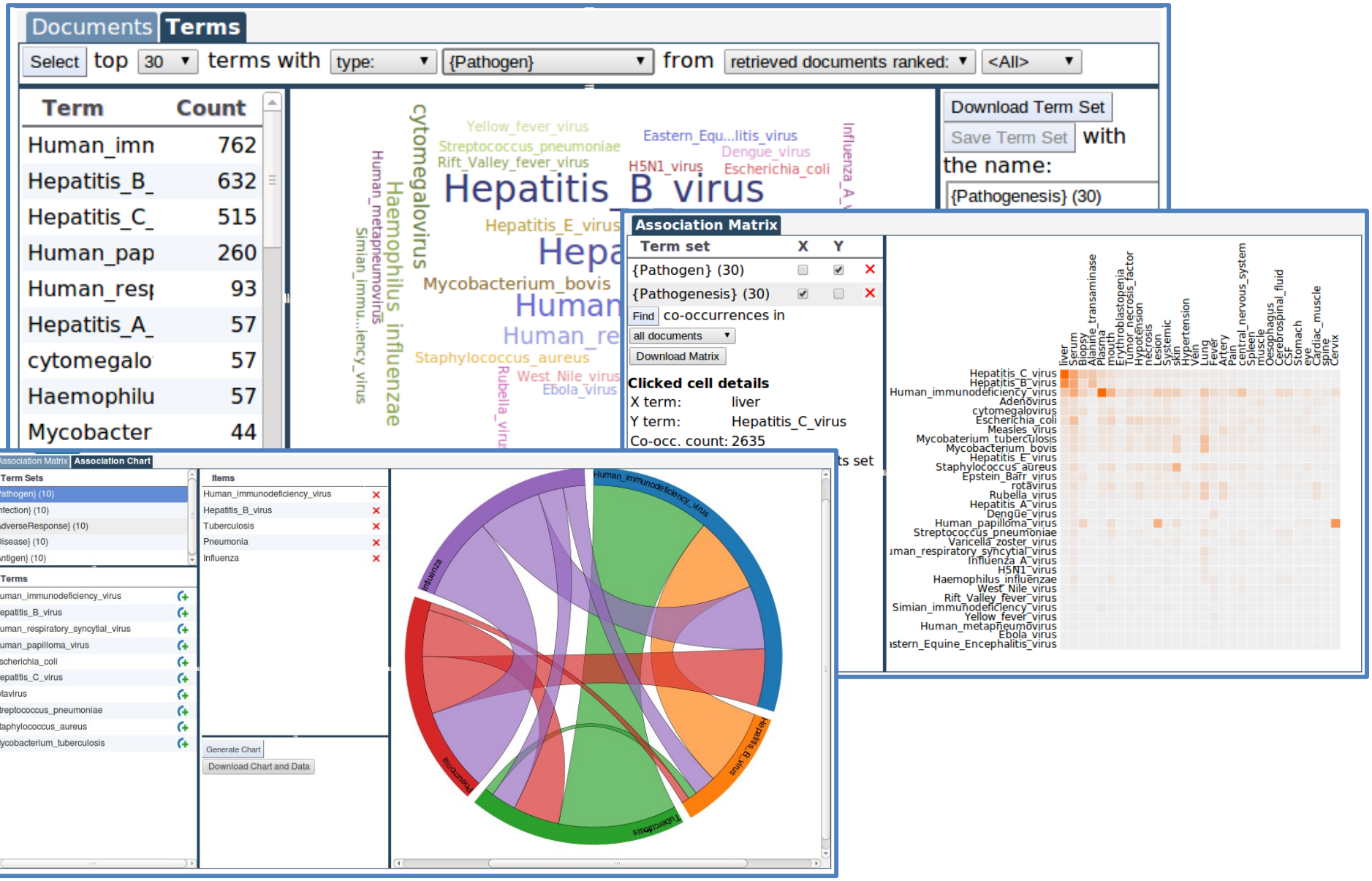
Prospector visualisations



Prospector visualisations



Prospector visualisations



Association Matrix

Term set	X	Y	
{Pathogen} (30)	☐	☑	✕
{Pathogenesis} (30)	☑	☐	✕

Find co-occurrences in

all documents ▾

Download Matrix

Clicked cell details

X term: liver

Y term: Hepatitis_C_virus

Co-occ. count: 2635

Show co-occurrence documents set in the Documents view.