
Module 16

Semantic Search

Module 16 schedule

9.45-11.00	<ul style="list-style-type: none">• xxx• Xxx
11.00-11.15	Coffee break
11.15-12.30	<ul style="list-style-type: none">• xxx• Xxx
12.30-14.00	Lunch Break
14.00-16.00	<ul style="list-style-type: none">• XXX• XXX

Module 16 outline

- Traditional approaches to search and retrieval
- Semantic annotation & search
- Overview of KIM and LifeSKIM platforms
- Demos

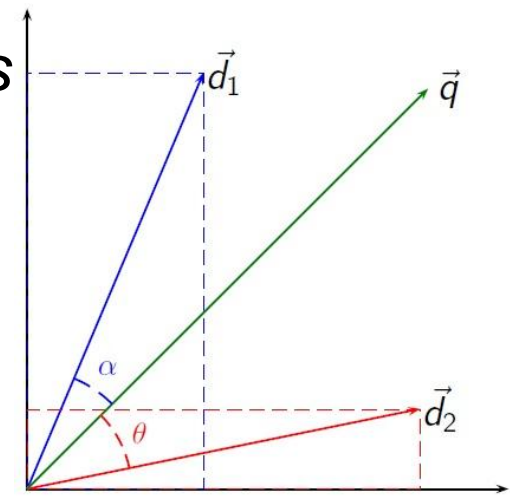
Traditional approaches to search and retrieval

IR models

- Boolean (set-theoretic)
 - Documents and queries are represented as sets (of terms/keywords)
 - Retrieval is based on set intersection
 - Advantages
 - Easy to implement
 - Disadvantages
 - Difficult to rank results
 - no term weighting

IR models (2)

- Algebraic
 - Documents and queries are represented as vectors in a multidimensional space (one dimension per term/keyword)
 - Retrieval is based on *vector similarities*
 - Cosine similarity
 - Advantages
 - Simple model
 - Ranking & Term weights
 - Disadvantages
 - Documents with similar topic but different vocabulary are not associated



Precision & Recall

- Precision
 - Measure of the quality of results
 - What % of the retrieved documents are relevant to the query?

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

- Recall
 - Measure of the completeness of results
 - What % of the documents which are relevant to the query are retrieved?

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Classical IR limitations

- Example
 - Query – “*Documents about a telecom companies in Europe related to John Smith from Q1 or Q2/2010*”
 - Document containing “*At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO*” will **not match**
 - Classical IR will fail to recognise that
 - Vodafone is a mobile operator, and mobile operator is a type of telecom
 - Vodafone is in the UK , which is part of Europe => Vodafone is a “telecom company in Europe”
 - 5th of May is in Q2 and John G. Smith may be the same as John Smith

Semantic Annotation & Search

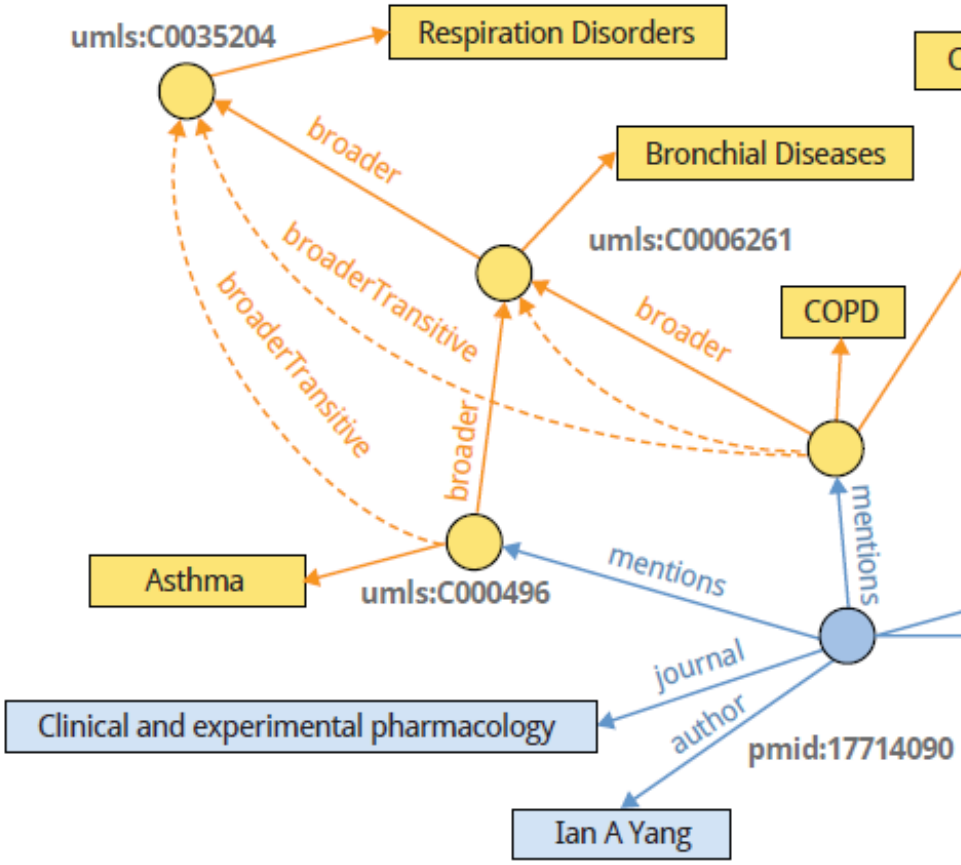
Semantic Annotation

- Semantic annotation (of text)
 - The process of linking text fragments to structured information
 - Organisations, Places, Products, Human Genes, Diseases, Drugs, etc.
 - Combines Text Mining (Information Extraction) with Semantic Technologies
- Benefits of semantic annotations
 - *Improves the text analysis process*
 - by employing Ontologies and knowledge from external Knowledge Bases / structured data sources

Semantic Annotation (2)

- Benefits of semantic annotations (cont.)
 - Provides *unambiguous (global) references for entities* discovered in text
 - Different from tagging
 - Provide the means for *semantic search*
 - Together or independently of the original text
 - Improved *data integration*
 - Documents from different data sources can share the same semantic concepts

Example

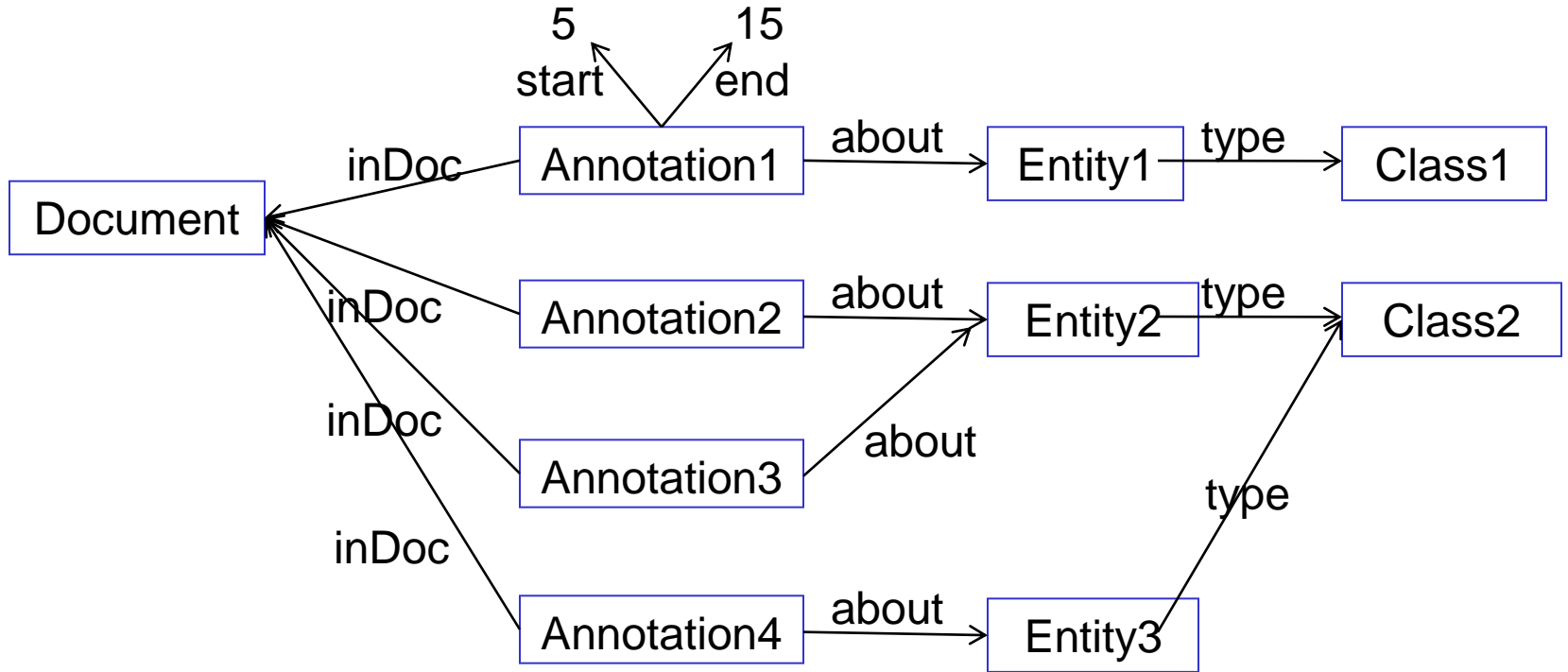


Asthma and chronic obstructive pulmonary disease (COPD) are chronic airway diseases characterized by airflow obstruction. The beta(2)-adrenoceptor mediates bronchodilatation in response to exogenous and endogenous beta-adrenoceptor agonists. Single nucleotide polymorphisms in the beta(2)-adrenoceptor gene (ADRB2) cause amino acid changes (e.g. Arg16Gly, Gln27Glu) that potentially alter receptor function.

Example (2)

- Demo of a GATE annotated document about “Asthma and chronic obstructive pulmonary disease”
 - Annotations of Genes
 - Each annotation is linked to an ontology *class*
 - Each annotation is linked to an ontology *instance*

Semantic Annotations



Semantic Search

- Semantic Search
 - In addition to the terms/keywords, explore the entity descriptions found in text
 - Make use of the semantic relations that exist between these entities
- Example
 - Query – “*Documents about a telecom companies in Europe related to John Smith from Q1 or Q2/2010*”
 - Document containing “*At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO*” will **not match**

Semantic Search (2)

- Classical IR will fail to recognise that
 - Vodafone is a mobile operator, and mobile operator is a type of telecom
 - Vodafone is in the UK , which is part of Europe
 - => Vodafone is a “telecom company in Europe”
 - 5th of May is in Q2
 - John G. Smith may be the same as John Smith

Types of Semantic Search

- What semantics?
 - Lexical semantics
 - Named entities
 - Factual knowledge
 - Ontologies / taxonomies
 - Hybrid approaches

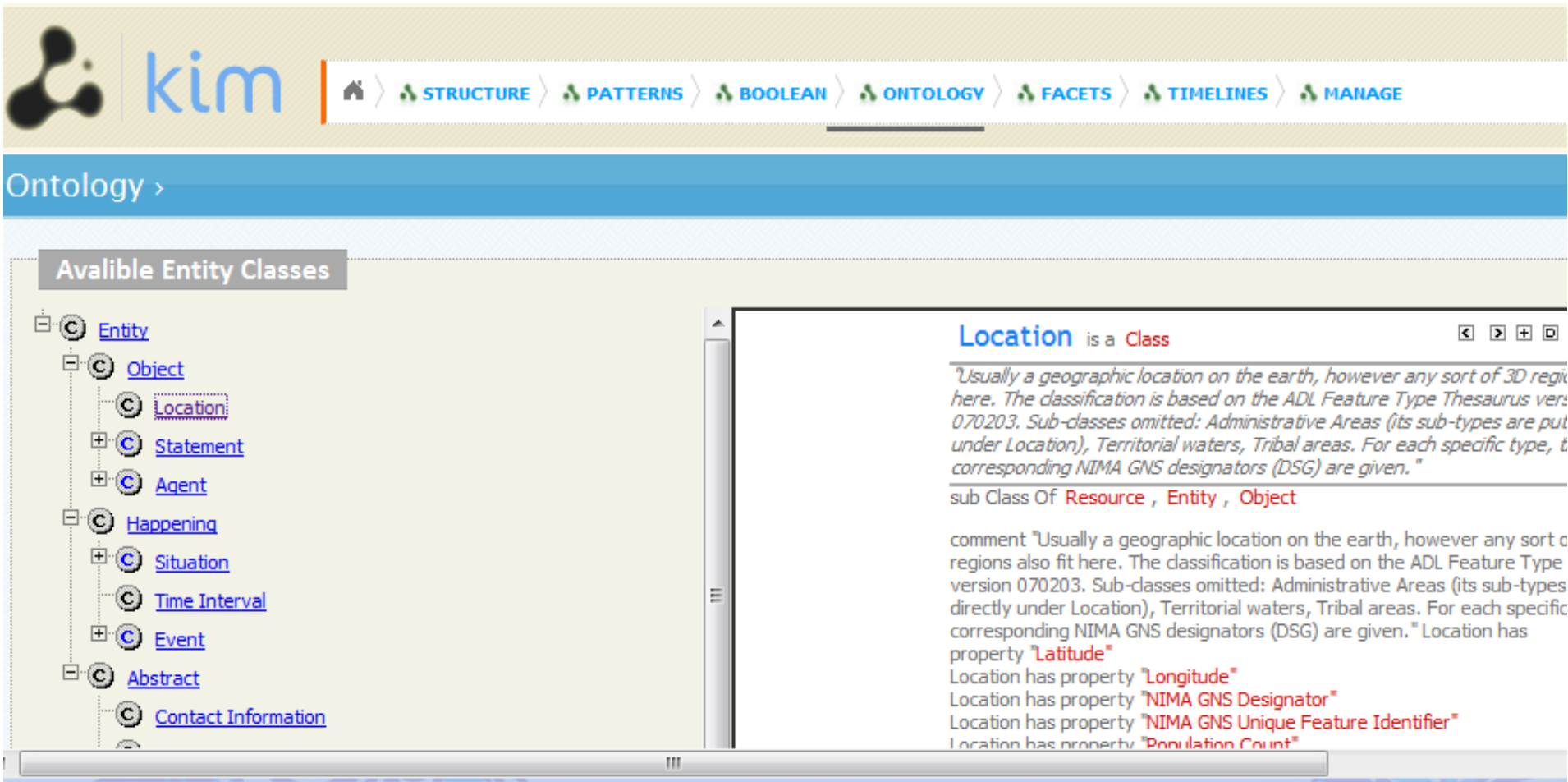
Types of Semantic Search (2)

- Types of queries
 - Occurrence
 - Co-occurrence
 - Structured queries
 - Faceted search
 - Pattern-matching

Types of Semantic Search (2)

- Structured queries
 - Query entities in the Knowledge Base
 - Very expressive and flexible
- Pattern queries
 - A set of predefined structured queries where *some* search criteria is already pre-specified
- Faceted search & navigation
 - Extracted entities are organised into *facets* (intelligent columns)
 - Easy to find documents that contain information about specific types of entities

Ontologies for semantic search



The screenshot shows the 'kim' ontology editor interface. At the top, there is a navigation bar with the following items: Home, STRUCTURE, PATTERNS, BOOLEAN, ONTOLOGY (highlighted), FACETS, TIMELINES, and MANAGE. Below the navigation bar is a blue header with the text 'Ontology >'. The main content area is titled 'Available Entity Classes' and displays a hierarchical tree of classes. The tree structure is as follows:

- Entity
 - Object
 - Location (highlighted)
 - Statement
 - Agent
 - Happening
 - Situation
 - Time Interval
 - Event
 - Abstract
 - Contact Information

The 'Location' class is selected, and its details are shown in a panel on the right. The details include:

- Location** is a **Class**
- Usually a geographic location on the earth, however any sort of 3D region here. The classification is based on the ADL Feature Type Thesaurus version 070203. Sub-classes omitted: Administrative Areas (its sub-types are put under Location), Territorial waters, Tribal areas. For each specific type, the corresponding NIMA GNS designators (DSG) are given.*
- sub Class Of **Resource** , **Entity** , **Object**
- comment "Usually a geographic location on the earth, however any sort of regions also fit here. The classification is based on the ADL Feature Type version 070203. Sub-classes omitted: Administrative Areas (its sub-types directly under Location), Territorial waters, Tribal areas. For each specific corresponding NIMA GNS designators (DSG) are given." Location has property **Latitude**
- Location has property **Longitude**
- Location has property **NIMA GNS Designator**
- Location has property **NIMA GNS Unique Feature Identifier**
- Location has property **Population Count**

Structured query in KIM

The screenshot shows the KIM (Knowledge In Memory) query interface. At the top, there is a navigation bar with the KIM logo and a breadcrumb trail: [STRUCTURE](#) > [PATTERNS](#) > [BOOLEAN](#) > [ONTOLOGY](#) > [FACETS](#) > [HYBRID](#) > [TIMELINES](#). Below this, a blue header bar contains the text "Structure >".

The main query area is titled "Lookup for patterns where". It contains three rows of query conditions:

- Row 1: "X, is a" with a dropdown menu showing "-----Person", "which name" with a dropdown menu showing "is unknown", and "and X" with a dropdown menu showing "hasPosition" and a variable "Y".
- Row 2: "Y, is a" with a dropdown menu showing "Job Position", "which name" with a dropdown menu showing "contains" and a text input field containing "spokesman", and "and Y" with a dropdown menu showing "withinOrganization" and a variable "Z".
- Row 3: "Z, is a" with a dropdown menu showing "Organization", "which name" with a dropdown menu showing "is exactly =", and a text input field containing "IBM".

To the right of the query area, a red-bordered box contains the text: *Show me all people who were mentioned as spokesmen in IBM*.

Below the query area is the "Attribute restrictions" section, which has two rows of dropdown menus for variables Y and Z, each with a "is unknown" dropdown.

Below that is the "Interested In" section, which has a dropdown menu showing "X, Y and Z".

At the bottom is the "Search for" section, which has two buttons: "DOCUMENTS" and "ENTITIES", and a "TRY FREE QUERY" button.

The browser's address bar and status bar are visible at the bottom of the screenshot.

Structured query example

- Demo of a structured query with KIM
 - Go to <http://ln.ontotext.com>
 - Select STRUCTURE
 - Build a query for:
 - *Persons* (unspecified name)
 - ... who have a *Position* of type *Job Position* (unspecified name)
 - ... within an *Organisation*
 - ... which is a *Company*
 - ... which name starts with “IBM”
 - Select
 - Entities
 - Documents mentioning the entities

Pattern query example (2)

- Demo of a structured query with KIM
 - Go to <http://ln.ontotext.com>
 - Select PATTERNS
 - Build a query for:
 - Organisations (unspecified name) located in Montreal
 - Select
 - Entities
 - Documents mentioning the entities

Faceted search in KIM

Facets >

Selected Items

- Montréal
- McGill University
- Researcher

Recent Items

- Quebec

People

25 of 28 shown below.

- Robert Sladek
- Philippe Froguel
- Frank Cirillo
- Sarah Gronberg Kolell
- David Ostry
- Sazzad Nasir
- Lara Pierce
- Erika Pearce
- Jim Tanaka
- Joel Gold
- Ian Gold
- Sahara Desert
- Markus Noethen
- Jim Carrey
- Hans Larsson
- Hans Dieter Sues
- Paul Sereno

Organizations

25 of 36 shown below.

- McGill University
- American Greetings
- University Of Electro...
- Department Of Telecommuni...
- Emerging
- The Associated Press
- Mit Media
- Michigan State University
- Osaka University
- Virtual University Design
- Canada Foundation For Inn...
- American Greetings
- National Academy of Scien...
- Imperial College
- Hallmark Cards, Inc.
- Hallmark Cards
- University Of Copenhagen

Locations

25 of 29 shown below.

- Montréal
- Tokyo
- Canada
- United States
- Kingdom of Denmark
- New Orleans
- City of London
- London
- Quebec
- Paris
- Philadelphia
- Kingdom of Morocco
- Federal Republic of Germa...
- Niger
- Republic of Niger
- Swiss Confederation
- Washington

Faceted search in KIM – document results



Document Keyword Filter

Matching documents: **13**
Between: 14 Feb, 2008 - 03 Sep, 2010

DOCUMENTS

TIMELINES

Documents, containing all selected entities

1-10 of 13 documents matching the search criteria.

Date	Title
06-09-2009	<p>Gene variant controls diabetic cells ...resistant to it. In Sunday's issue of the journal Nature Genetics, researchers from Imperial College London and Copenhagen in Denmark reported genetic variants... ...in establishing [Type 2 diabetes] risk," Dr. Robert Sladek of McGill and his co-authors concluded. The re associated with ...</p>
19-11-2009	<p>3 new ancient crocodile species fossils found ...- like wild boar tusks - roamed parts of northern Africa millions of years ago, researchers reported Thursday. The fossils... ...detailed Thursday by researchers Paul Sereno of the University of Chicago and Hans Larsson of McGill University, which ... National Geographic Society, which ...</p>
03-08-2009	<p>Future tech on show at 36th SIGGRAPH ...ORLEANS (AP) -- If you pull on my ear, will I follow you anywhere? Yes, say researchers at University of Illinois at Urbana-Champaign directed from... ...new drugs and new buildings, Haley said. But immersive VR is still in the works. At McGill University in Montreal, researchers are testing a device that by vibrating ...</p>
06-09-2009	<p>Gene variant controls diabetic cells ...resistant to it. In Sunday's issue of the journal Nature Genetics, researchers from Imperial College London and Copenhagen in Denmark reported genetic variants... ...in establishing [Type 2 diabetes] risk," Dr. Robert Sladek of McGill and his co-authors concluded. The re associated with ...</p>
03-11-2009	<p>Learning to talk changes how we hear speech: study ...News The robotic device used in the language experiment to isolate the movements involved in talking, determined that learning to talk changes...</p>

Faceted search example

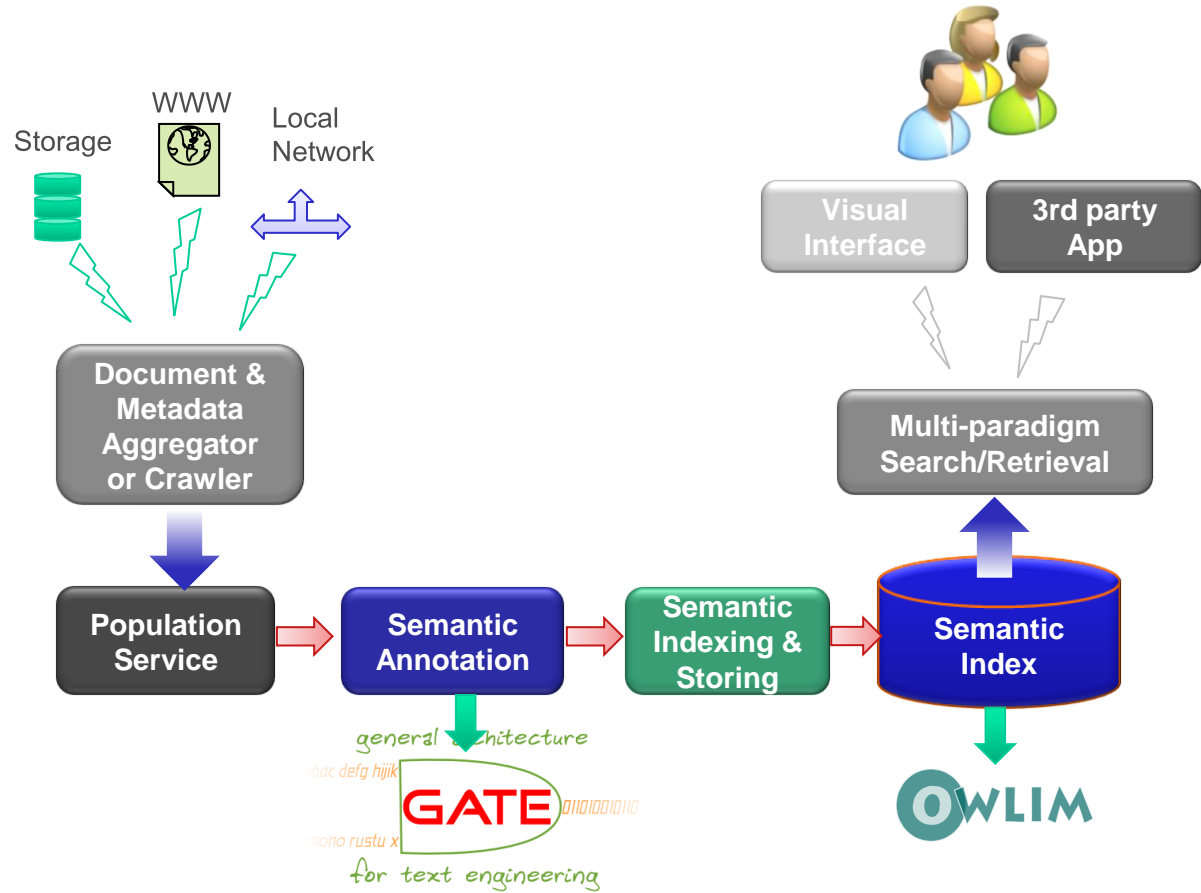
- Demo of a faceted navigation with KIM
 - Go to <http://ln.ontotext.com>
 - Select “Facets”
 - Restrict “Organisations” to “McGill University”
 - Restrict “Locations” to “Montreal”
 - Select “researcher” from “Related Entities”
 - (document results displayed on bottom of page)

Overview of KIM and LifeSKIM

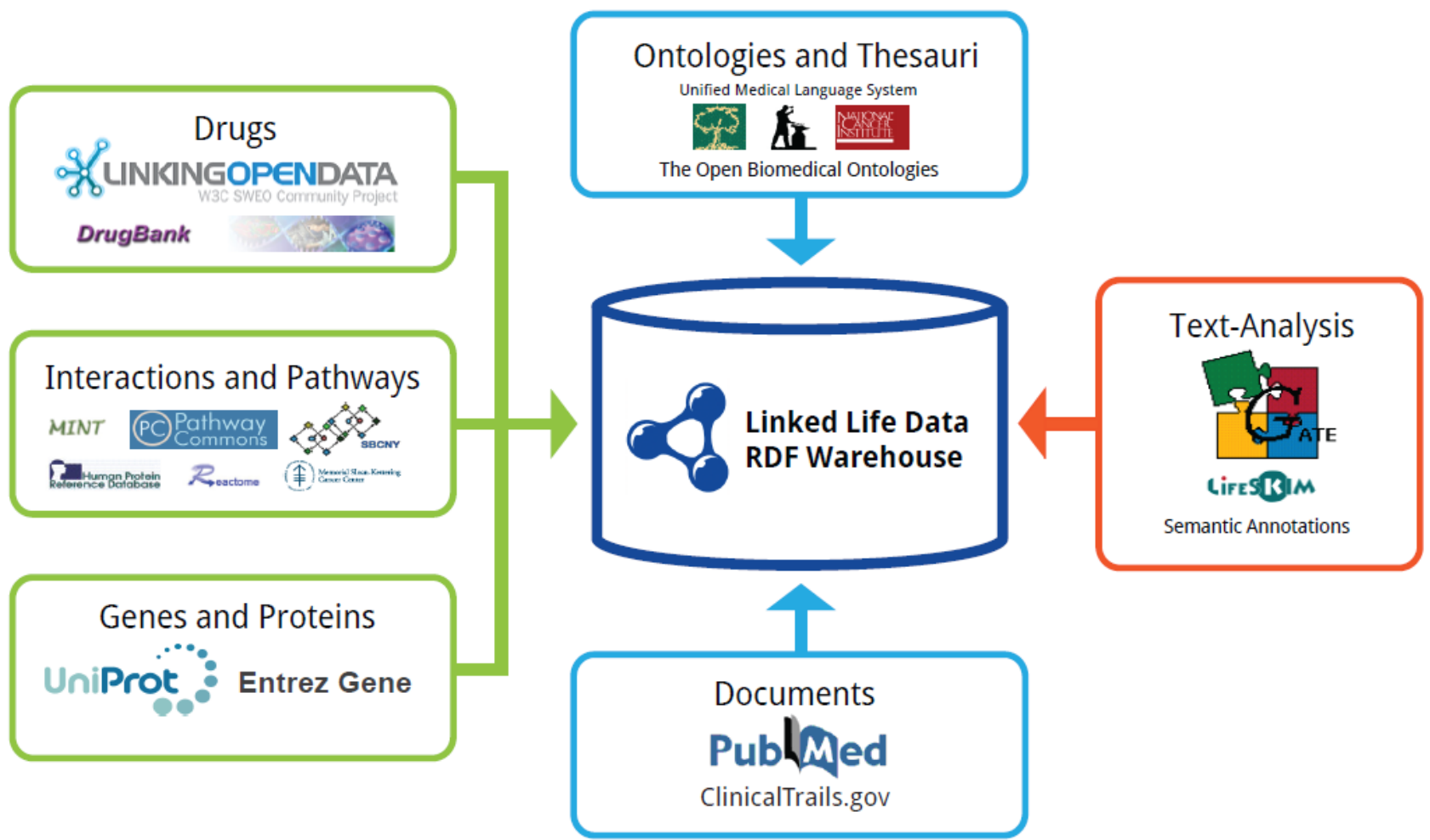
The KIM Platform

- A platform offering services and infrastructure for:
 - **Automatic semantic annotation** of text
 - **Text-mining** and ontology population
 - **Semantic indexing and retrieval** of content
 - Query and navigation across heterogeneous text and data
- Based on an Information Extraction technology
 - built on top of GATE
- Offers unparalleled **heterogeneous querying** facilities

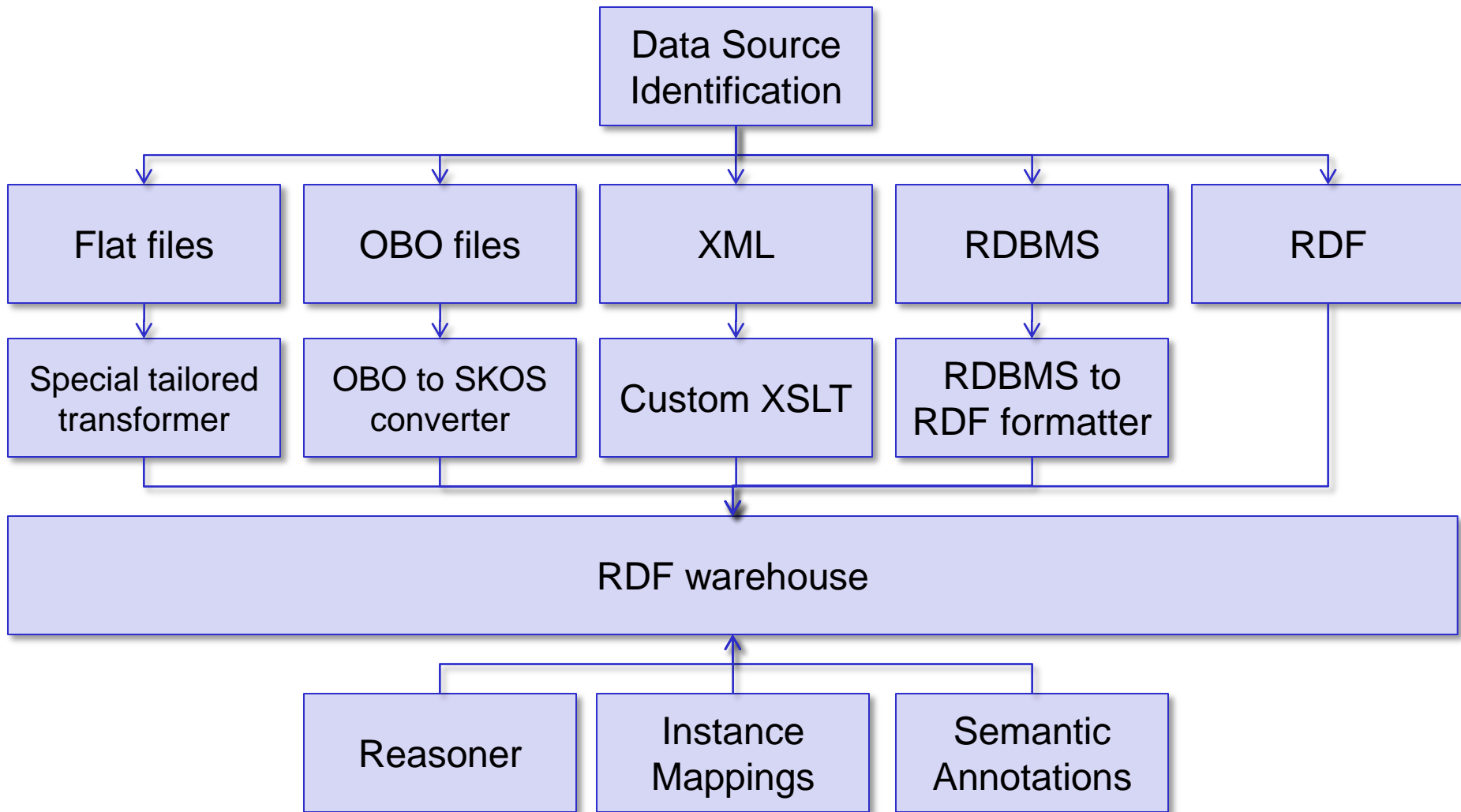
KIM platform (2)



LifeSKIM & Linked Data



LifeSKIM / Linked Data ETL



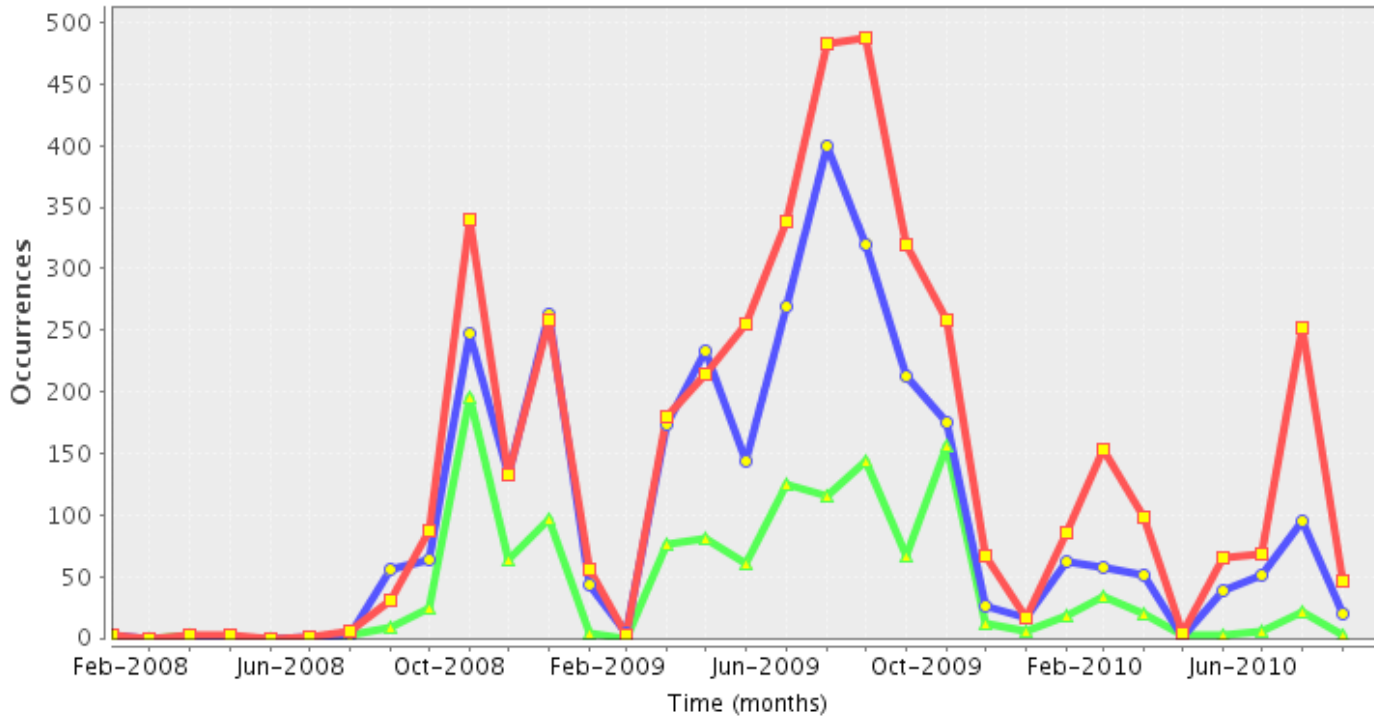
Timelines for entity popularity in KIM

- Timelines for entity occurrences over some period of time
 - Can be used & extended for sentiment analysis

Timelines in KIM

Timelines for entities:

Google Inc., Microsoft Corporation, Yahoo! Inc.



- Google Inc. (4313)
- Microsoft Corporation (3162)
- Yahoo! Inc. (1347)

Timelines example

- Demo of timeline with KIM
 - Go to <http://ln.ontotext.com>
 - Select “Timelines”
 - Build a monthly timeline comparing mentions of Concordia, McGill and University of Montreal
 - Time period: max
 - Granularity: month
 - Based on: occurrences