



K I M

Knowledge and Information
Management Platform

KIM Platform

An Overview

(c) Copyright 2002-2006 Ontotext Lab, Sirma Group Corp.

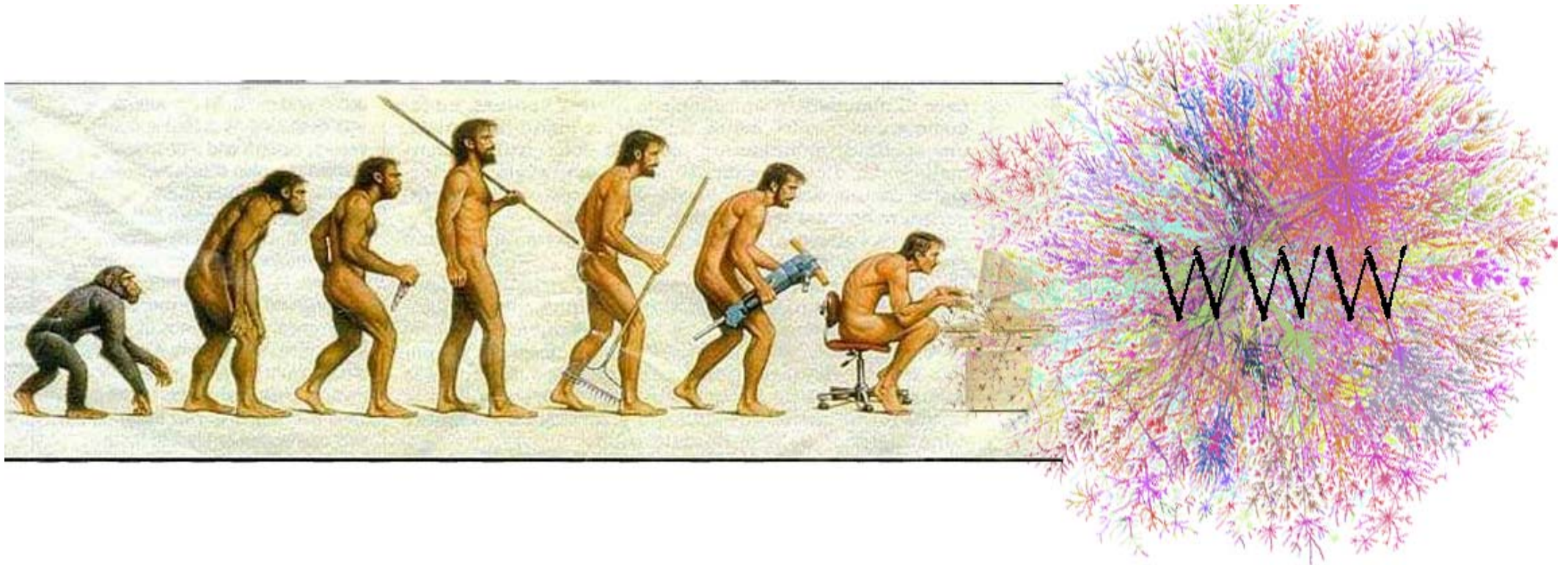
30 Sep, 2006

Presentation Outline

- **What: Functionality**
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- **How: Architecture & Implementation**
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- **Wrap up**

Why?

Instead of blah-blah about the **information overload** and the biggest library created by the human kind ...



Presentation Outline

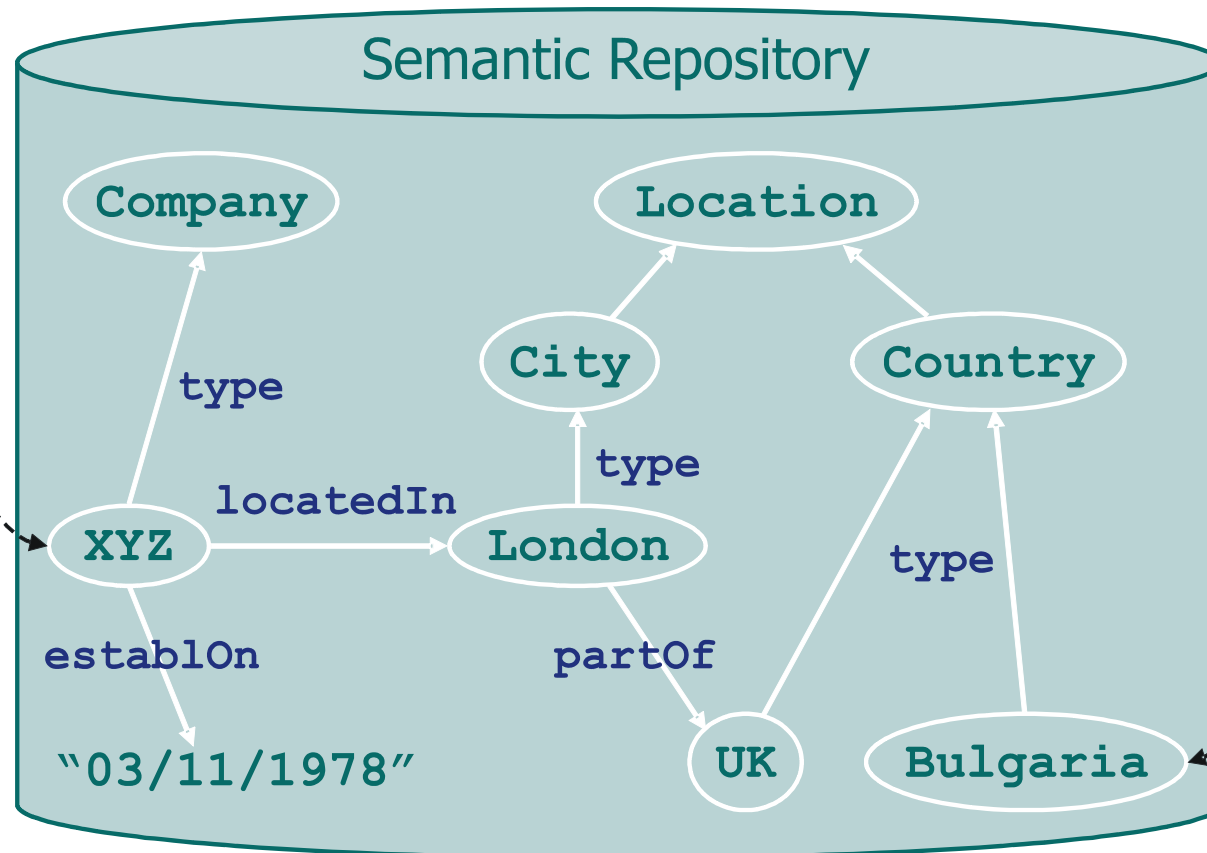
- What: Functionality
 - **Text-Mining, Semantic Annotation, and Hyper-linking**
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- Wrap up

Semantic Annotation, Indexing, and Retrieval

- A platform offering software infrastructure for:
 - (semi-)automatic **semantic annotation** of text
 - **ontology population**
 - Store the extracted facts and reason on top of them
 - **semantic indexing** and **retrieval** of content
 - **query** and **navigation** involving **structured knowledge**
- Based on **Information Extraction** (i.e. text-mining) technology
- It was designed to **enable Semantic Web** applications ...
 - by providing a **metadata generation** technology
 - in a **standard, consistent, and scalable framework**
- But appeared suitable for **Knowledge Management and BI**

What KIM does? - Semantic Annotation

XYZ announced profits in Q3, planning to build a \$120M plant in Bulgaria, and more and more and more and more text..



Simple Usage: Highlight, Hyperlink, and...

Guardian Unlimited | World Latest | EPA Moving on New Front to Cut Pollution - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.guardian.co.uk/uslatest/story/0,1282,-4076961,00.html> Go

KIM Plugin

Annotate Clear About

Entity

- Abstract
- Happening
 - Event
 - Situation
 - TimeInterval
- Object
 - Agent
 - Organization
 - Person
 - BusinessObject
 - InformationResource
 - Location
 - Statement
 - Vehicle

Classes Entities Config

Place Links

Breaking news US

EPA Moving on New Front to Cut Pollution

Tuesday May 11, 2004 7:46 AM

By H. JOSEF [HEBERT](#)

[Associated Press](#) Writer

[WASHINGTON](#) ([AP](#)) - The government is moving on a new front to cut air pollution. This time ferry boats and harbor tugs, farm tractors and train locomotives, and dirt movers at construction sites are the targets.

[Cheney to Have Routine Check of Pacemaker](#)
8:46 am

[GOP Seeks to End Tax Cut Debate](#)
8:46 am

[Govt. Grounds](#)

The [Environmental Protection Agency](#) is issuing new regulations aimed at cutting the amount of smog-causing chemicals and fine soot that comes from these off-road diesel-powered vehicles

Internet

Simple Usage: ... Explore and Navigate

The image shows two overlapping browser windows. The left window, titled 'KIM Explorer - Microsoft Internet Explorer', displays a semantic annotation interface for 'The Associated Press, a NewsAgency, Trusted'. It features a table of properties and values, and a section for related entities.

Property	Value
hasAlias	The Associated Press
hasAlias	AP
hasAlias	Associated Press
locatedIn	New York
locatedIn	New York

Resource	Link to The Associated Press
correspondent	withinOrganization
correspondent	withinOrganization
correspondent	withinOrganization
correspondent	withinOrganization

Copyright © 2004 Ontotext Lab, Sirma AI, Bulgaria

The right window, titled '... - Microsoft Internet Explorer', shows a news article with semantic annotations. A large orange arrow points from the KIM Explorer window to the article. The article text includes: 'May 11, 2004 7:46 AM', 'SEF HEBERT', 'ed Press W', 'STON (AP) - ... government is moving on a new front to cut air pollution. This time ferry boats and harbor tugs, farm tractors and train locomotives, and dirt movers at construction sites are the targets.', 'The Environmental Protection Agency is issuing new regulations aimed at cutting the amount of smog-causing chemicals and fine soot that comes from these off-road diesel-powered vehicles', and 'Govt. Grounds'.

Classes: BusinessObject, InformationResource, Location, Statement, Vehicle

Place Links:

Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - **Co-occurrence and Popularity Timelines**
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- Wrap up

CORE: Co-occurrence and Ranking of Entities

Be able to efficiently query for:

- **Number of appearances** and **popularity** of entities

Q1: How often has a company appeared in the international business news during a given period ?

- **Co-occurrence** of entities

Q2: Give me the people that co-appear with telecom companies

- Combination of the above with **semantic queries** and **Full-Text Search**, time-constraints, etc.

Q3: Q2 + where the documents from 2004 contain "fraud" and the company is located in South-east Europe

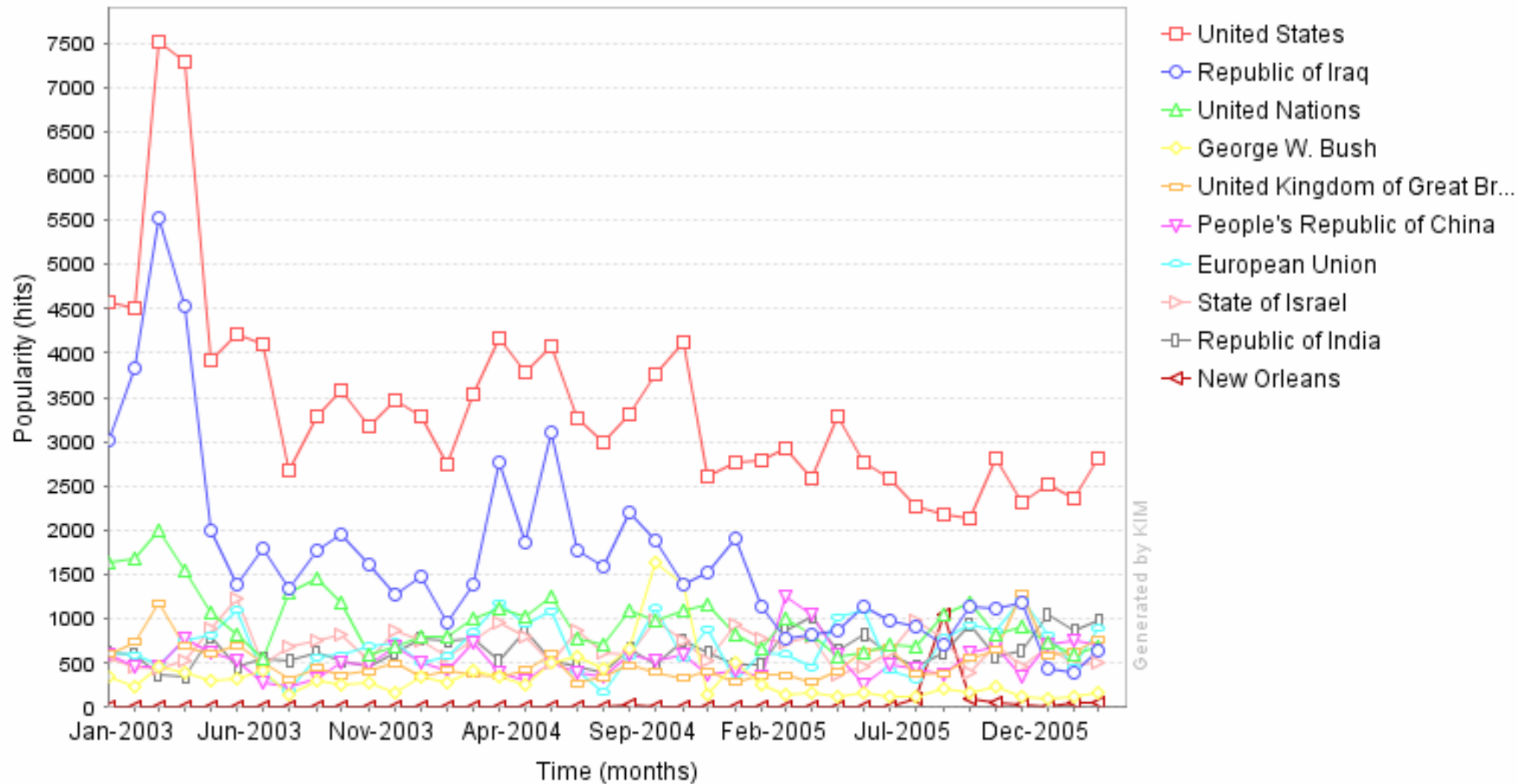
- **Popularity ranking**

Q4: the 5 most popular persons for each month in 2005, based on news for South Africa, showing a timeline of their ranking

CORE: Scale and Applications

- Allow such queries in *efficient* manner over data with cardinality:
 - 10^6 entities/terms in 10^7 documents (tens of millions)
 - 10^2 entities occurring in an average document
 - managing and querying efficiently 10^9 entity occurrences!
- **Detection of “associative” links** between entities
 - based on co-occurrence in context;
 - an alternative to extraction of “strong links” by parsing local context
- **Media monitoring**: the ranking is as good/relevant/representative as the set of documents is
- **Computing timelines** for entity ranking or co-occurrence
 - “How did our popularity in the IT press changed during June” (i.e. “What is the effect of this 1.5MEuro media campaign ?!?”)
 - “How does the strength of association between organization X and RDF changes over Q1 ?”

Timelines for Most popular Entities




Document Filter: ALL docs, containing Keywords : (none) Entities : (none)

Time Period: 01/01/2003 to 31/03/2006 Granularity: Month

Options: display 2 topmost entities of type Entity for each time unit




CORE Search

[Clear](#) | [Options](#)



- > Home
- > Entity Pattern Search
- > Predefined Patterns
- > Entity Lookup
- > Keyword Search
- > Browse Ontology
- > CORE Search
- > Timelines
- > About KIM

Powered by:

CORE Search

Document Keyword Filter

Matching documents: **63963**

Documents
Timelines

Selected Items Filter

(No items selected)

Recent Items

(No recent items)

Key Phrases

25 of **14488** shown below

- abortion
- bloc
- chair
- church
- crowd
- earthquake
- election
- Faith
- freedom
- influence
- island
- Mass
- member nation
- papacy
- people
- Pope
- post
- quake
- reign
- scale
- sense
- summit
- trip
- Visit
- World

People

25 of **63161** shown below

- A. Tarzartes
- Bruce L.A. Carter
- Craig D. Shimasaki
- D. Gregory Smith
- Frank Chang
- Fred Hiller
- Fritz Horlacher
- George B. Rathmann
- Horace J. Davis III
- Ian Kindred
- J. Bruce Robinson
- James A. Johnson
- James J. Jim Schiro
- James R. Tolbert III
- Julie A. Sansom-Reese
- Lodewijk Christiaan van
- Makoto Kaneko
- Markus Haefeli
- Norman R. Proulx
- Patrick O'Sullivan
- Paul F. Forman
- Peter Eckert
- Richard M. Dressler
- Roberto R. Romulo
- William Kim Wah Tan

Organizations

25 of **27622** shown below

- Amhrest College
- Brandeis University
- Brigham Young Universit
- Drexel University
- Nanyang Technological I
- New Mexico State Unive
- Queen Marry University
- Sofia University
- Stanford University
- Technical University Kos
- TU Vienna
- University of Arizona
- University of Economics,
- University of Georgia
- University of Maryland
- University of Maryland C
- University of Santiago de
- University of South Calif
- University of Southern C
- University of Strathclyde
- University of Twente
- University of Washingtor
- Vassar College
- Vrije Universiteit Amsterc
- Vrije Universiteit, Brusse

Locations

25 of **21568** shown below

- Bandirma Korfezi
- Bristol Channel
- Dzharylhats'ka Zatoka
- Erdek Korfezi
- Feodosiys'ka Zatoka
- Garabogazkol Aylagy
- Gemlik Korfezi
- Golfo de Venezuela
- Golfo Triste
- Gulf of Aden
- Gulf of Antalya
- Gulf of Iskenderun
- Gulf of Paria
- Ildir Korfezi
- Izmit Korfezi
- Karkinit's'ka Zatoka
- Kusadasi Korfezi
- Persian Gulf
- Saros Korfezi
- Taganrogskiy Zaliv
- Tendrivs'ka Zatoka
- Turkmenbasy Aylagy
- Yahorlyts'kyy Lyman
- Zaliv Adzhibay
- Zatoka Syvash

Copyright © 2005 Ontotext Lab, Sima Group Corp.

Name Restriction

CORE Search

| Clear | Options |

Document Keyword Filter	Key Phrases	People	Organizations	Locations
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="Austr"/>
	25 of 14488 shown below	25 of 63161 shown below	25 of 27622 shown below	25 of 34 shown below
Matching documents: 63963 <input type="button" value="Documents"/> <input type="button" value="Timelines"/>	abortion bloc chair church crowd earthquake election Faith freedom influence island Mass member nation papacy people Pope post quake reign scale sense summit trip Visit World	A. Tarzartes Bruce L.A. Carter Craig D. Shimasaki D. Gregory Smith Frank Chang Fred Hiller Fritz Horlacher George B. Rathmann Horace J. Davis III Ian Kindred J. Bruce Robinson James A. Johnson James J. Jim Schiro James R. Tolbert III Julie A. Sansom-Reese Lodewijk Christiaan van Makoto Kaneko Markus Haefeli Norman R. Proux Patrick O'Sullivan Paul F. Forman Peter Eckert Richard M. Dressler Roberto R. Romulo William Kim Wah Tan	Amhrest College Brandeis University Brigham Young Universit Drexel University Nanyang Technological U New Mexico State Unive Queen Marry University Sofia University Stanford University Technical University Kos TU Vienna University of Arizona University of Economics, University of Georgia University of Maryland University of Maryland C University of Santiago de University of South Calif University of Southern C University of Strathclyde University of Twente University of Washingtor Vassar College Vrije Universiteit Amsterc Vrije Universiteit, Brusse	Austr Australia Zoo Australian Capital Territo Australian High Tech Crib Austrre Porsangneset central Australia Commonwealth of Austr eastern Australia Lower Austria north Australia North Western Australia north-east Australia north-eastern Australia northern Australia Republic of Austria South Australia south-east south-east southern Austria State of South Australia State of Western Australia Territoire des Terres Aus Upper Austria Western Australia western Australia Western Australia Mount
Selected Items Filter (No items selected)				<input type="text" value="Republic of Austria"/>
Recent Items <input type="button" value="+"/> Republic of Austria				

Co-occurring Entities

CORE Search

| Clear | Options |

Document Keyword Filter	Key Phrases	People	Organizations	Locations
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
	25 of 42 shown below	Loading ...	19 matching entities	25 of 49 shown below
Matching documents: 13 <input type="button" value="Documents"/> <input type="button" value="Timelines"/>	Bird case city coalition death governor honour judge name news agency newspaper paper parliament picture reform republic reputation resident ring son spokesman strike town union virus	Adolf Hitler Archbishop Rowan Willie Arnold Schwarzenegger Axel Springer Verlag Benita Brian L Christian Fraser Christopher D. Hughes Deborah Lipstadt Donald Beardslee Greville Janner Irving Joerg Haider Markos Kyprianou Matthew Cooper Mozart Otto Schneider Pilz Quentin State Prison Rudolf Gollia Samuel Jutzi Siegfried Nagl Stanley Stephen Smith	AFP Austrian TV BBC BBC News Court of Appeals Green Party Holocaust Educational Tr Mediaprint National Federation of Au Reuters Group PLC Robert Koch Institute Supreme Court of Pakista UK United Nations US and British embassy US Army US Circuit Court of Appe	Brussels Bundesland Karnten Bundesland Salzburg Bundesland Steiermark Bundesland Wien California City of London Douglas Hollywood Innsbruck Kingdom of Spain Kingdom of Sweden Kingdom of the Netherlar Linz London Los Angeles Poland Portuguese Republic Rome Sacramento Split U K United States Vienna
Selected Items Filter <input type="checkbox"/> Graz <input type="checkbox"/> Republic of Austria				
Recent Items <input type="checkbox"/> Republic of Austria				

Co-occurrence...execution

CORE Search

| Clear | Options |

Document Keyword Filter	Key Phrases	People	Organizations	Locations
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Matching documents: 5 <input type="button" value="Documents"/> <input type="button" value="Timelines"/>	20 matching entities case city death <input type="checkbox"/> Execution governor honour injection innocence name news agency picture republic reputation request resident ring son sport stadium sun victory	21 matching entities Archbishop Rowan Willie Benita Ferrero- Waldner Christopher D. Hughes Denzel Washington Dieter Hardt-Stremayr Donald Beardslee Joerg Haider Karl Kling Matthew Cooper Monika Ficszko Mozart Peggy Ryen Pilz Quentin State Prison Sean Penn Siegfried Nagl Stanley Thomas Rajakovics Waltraud Klasnic Wolfgang Schuessel	8 matching entities AFP Austrian TV Court of Appeals Green Party Reuters Group PLC Supreme Court of Pakist US and British embassy US Circuit Court of Appe	15 matching entities Bundesland Karnten Bundesland Steiermark Bundesland Wien California Commonwealth of Austr Douglas Europe Hollywood Liebenau Stadium Los Angeles Sacramento Schwarzenegger Stadiu United States
Selected Items Filter <input type="checkbox"/> Arnold Schwarzenegger <input type="checkbox"/> Graz <input type="checkbox"/> Republic of Austria				
Recent Items <input type="checkbox"/> Republic of Austria				

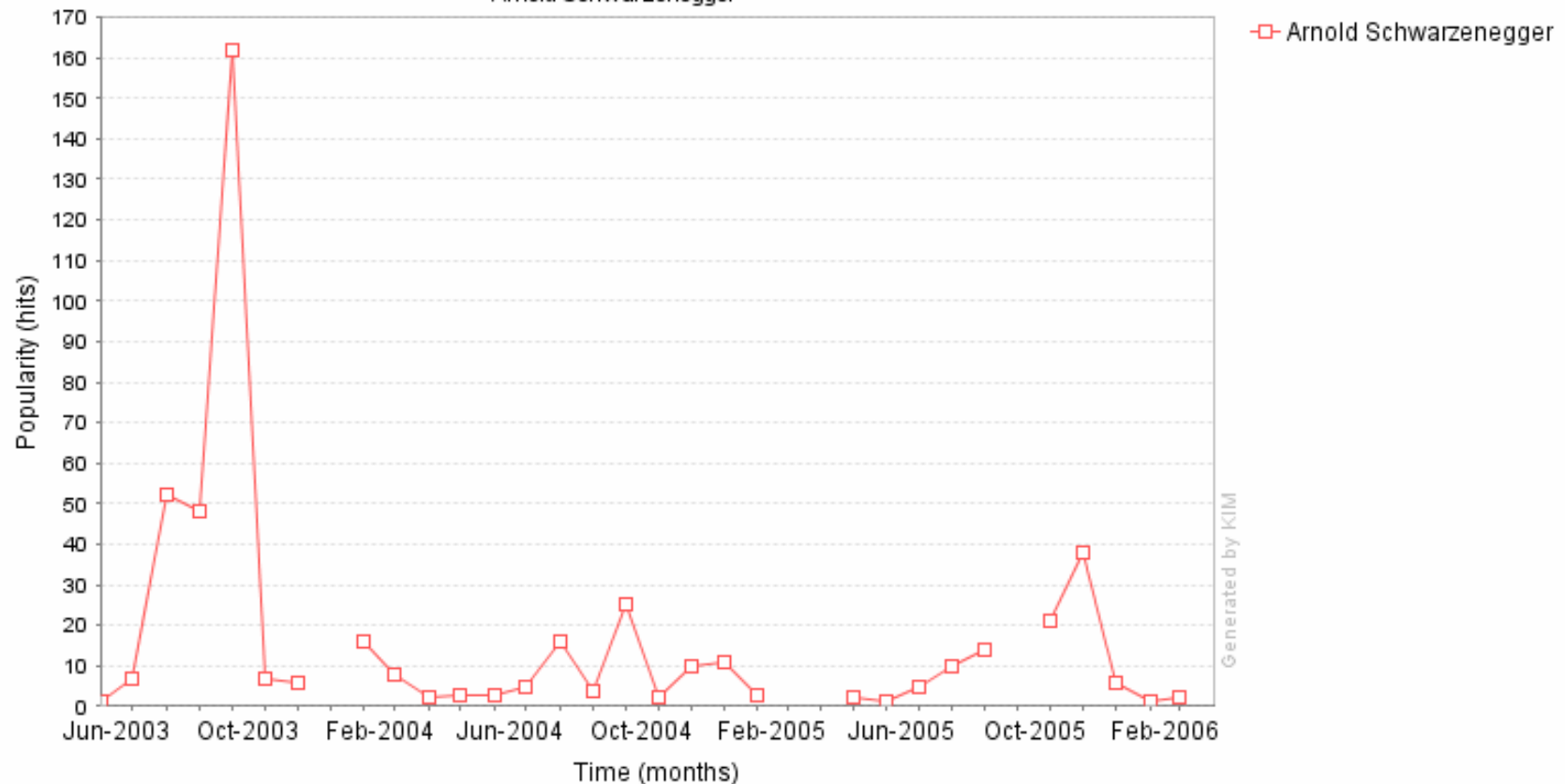
Search Form

Arnold's Popularity

Timelines Result

Timelines for entities :

Arnold Schwarzenegger



Document Filter: 167 docs, containing Keywords : (none) Entities : Arnold Schwarzenegger

Time Period: 01/01/2003 to 31/03/2006 Granularity: Month

The Documents, Forming the Peak

Document Query Result

Date	Title
07/10/2003 21:46	Q&A: California recall vote
06/10/2003 08:06	Profile: Cruz Bustamante
29/10/2003 04:31	Forecast brings hope for stricken state
29/10/2003 20:17	Firefighter dies in California blaze
09/10/2003 06:40	Arnie steals Ugandan thunder
02/10/2003 20:15	Sex scandal draws Arnie apology
03/10/2003 09:47	Gropes not arms interest US
07/10/2003 21:47	Californians make up their minds
08/10/2003 03:43	Schwarzenegger win dominates US media
09/10/2003 04:14	Maria Shriver: Arnie's secret weapon
09/10/2003 04:20	World press digests Arnie's latest role
07/10/2003 21:33	US diary: California recall
08/10/2003 04:08	Arnie wins California election
08/10/2003 02:02	Arnie win 'sign of US discontent'
10/10/2003 02:49	Arnie names transition team

1-15 of 32 Documents per page:

Available Query Options

Timelines

Document Distribution



Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - **Combining FTS, Structured Queries, and Inference**
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- Wrap up

How KIM Searches Better

KIM can match a **Query**:

Documents about a telecom company in Europe, John Smith, and a date in the first half of 2002.

With a document containing:

At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO

The classical IR could not match:

- Vodafone with a "telecom in Europe", because:
 - Vodafone is a mobile operator, which is a sort of a telecom;
 - Vodafone is in the UK, which is a part of Europe.
- 5th of May with a "date in first half of 2002";
- "John G. Smith" with "John Smith".

Entity Pattern Search

The screenshot shows a web browser window titled "KIM WEB UI - Microsoft Internet Explorer" with the address bar containing "http://ontotest.sirma.bg/KIM/screen/EntityPatternSearch.jsp". The main content area is titled "Pattern Search" and features a sidebar on the left with a navigation menu: "Datastore", "Entity Pattern Search", "Predefined Patterns", "Entity Lookup", "Keyword Search", and "About KIM".

The search interface is divided into several sections:

- Lookup for patterns where:**
 - X, is a , which name
 - and X Y
 - Y, is a , which name
 - and Z
 - Z, is a , which name
- attribute restrictions:**
 -
 -
 -
- Interested In:**
 -
- Search for:**
 -

At the bottom left, there are logos for "Powered by:" including GATE, RDF SESAME, and Lucene.

Pattern Search: Entity Results

The screenshot shows a web browser window titled "KIM WEB UI - Microsoft Internet Explorer". The address bar contains the URL "http://ontotest.sirma.bg/KIM/screen/EntitiesResult.jsp". The page content is divided into a left sidebar and a main content area.

Left Sidebar:

- KIM logo
- Navigation menu:
 - Datastore
 - Entity Pattern Search
 - Predefined Patterns
 - Entity Lookup
 - Keyword Search
 - About KIM
- Powered by:
 - GATE logo
 - RDF SESAME logo
 - Lucene logo

Main Content Area: Entity Query Result

Company	Type	Tr
<input type="checkbox"/> Groupe Danone	PublicCompany	+
<input type="checkbox"/> Groupe Danone World Water Division	Company	+
<input type="checkbox"/> Labeyrie	Company	+
<input type="checkbox"/> Groupe Lactalis	Company	+
<input type="checkbox"/> Taittinger S.A.	Company	+
<input type="checkbox"/> IAWS Group plc	PublicCompany	+
<input type="checkbox"/> Quilmes Industrial S.A.	PublicCompany	+
<input type="checkbox"/> Cadbury Schweppes Beverage Unit	Company	+
<input type="checkbox"/> Cadbury Schweppes plc	PublicCompany	+
<input type="checkbox"/> Dairy Crest Group plc	PublicCompany	+
<input type="checkbox"/> Gallaher Group Plc	PublicCompany	+
<input type="checkbox"/> Halewood International Limited	Company	+
<input type="checkbox"/> Ranks Hovis McDougall Limited	Company	+
<input type="checkbox"/> SABMiller plc	PublicCompany	+
<input type="checkbox"/> Tate & Lyle PLC	PublicCompany	+

1-15 of 16 Entities per page: 15 [v] [Next] [Last]

Available Query Options

[Refine Query] [New Query] [Edit Query]

Entity Pattern Search: KIM Explorer

The screenshot displays two overlapping browser windows. The background window, titled 'KIM WEB UI - Microsoft Internet Explorer', shows a search interface with a menu on the left and a list of search results. The foreground window, titled 'KIM Explorer - Microsoft Internet Explorer', displays a detailed view of the selected entity, 'Groupe Danone, a PublicCompany, Trusted^{tip!}'. This view includes a table of properties and values, a comment, and other metadata.

KIM WEB UI - Microsoft Internet Explorer

Address: <http://ontotest.sirma.bg/KIM/screen/EntitiesResult.jsp>

Entity Query

- Company
 - Groupe Danone
 - Groupe Danone
 - Labeyrie
 - Groupe Lactalis
 - Taittinger S.A.
 - IAWS Group plc
 - Quilmes Industria
 - Cadbury Schwep
 - Cadbury Schwep
 - Dairy Crest Grou
 - Gallaher Group P
 - Halewood Intern
 - Ranks Hovis McD
 - SABMiller plc
 - Tate & Lyle PLC

1-15 of 16 Entities

Available Query Options

Refine Query

Powered by:

-
-
-

KIM Explorer - Microsoft Internet Explorer

Groupe Danone, a PublicCompany, Trusted^{tip!}

Property	Value
hasMainAlias	Groupe Danone
comment	"You say Danone, I say Dannon; let's call the whole thing one of the largest food producers in the world. Groupe Danone is the global leader in cultured dairy products (including yogurt, cheese, and dairy desserts) and biscuits (cookies, crackers, and snacks). Its Evian and other brands make it #2 in bottled water (behind Nestlé). Danone has dozens of regional and international brands, including Dannon yogurt (US), Jacob's and LU cookies and crackers, and HP and Lea & Perrins sauces. It owns almost 45% of BSN Emballage, a maker of glass containers. More..."
activeInSector	Food, Beverage & Tobacco
hasWebPage	http://www.danonegroup.com
locatedIn	French Republic
tradedOn	New York Stock Exchange
stockExchangeIndex	"DA"
FISCAL_SALES	"\$12,897 mln."
FISCAL_NET_INCOME	"\$118 mln."
numberOfEmployees	"100,560"
fullyOwns	Danone

Graph Knowledge Explorer

The screenshot displays a web browser window titled "TouchGraph GraphLayout - Microsoft Internet Explorer". The address bar shows the URL: `http://62.213.161.156/KIM/graph/Graph.jsp?uri=http://www.ontotext.com/kim/kimo.rdfs%23PublicCompany_T.138`. The main content area features a central node labeled "Groupe Danone" with the description "a trusted PublicCompany". This central node is connected to several other nodes via labeled relationships:

- Jacques Vincent** (holder) - **hasPosition** - **Vice Chairman and COO**
- Franck Riboud** (holder) - **hasPosition** - **Chairman and CEO**
- Food, Beverage & Tobacco** - **activeInSector** - **Groupe Danone**
- French Republic** - **locatedIn** - **Groupe Danone**
- Danone** - **fullyOwns** - **Groupe Danone**
- http://www.danonegroup.com** - **hasWebPage** - **Groupe Danone**
- New York Stock Exchange** - **tradedOn** - **Groupe Danone**
- EVP, Finance** - **withinOrganization** - **Groupe Danone**
- Chairman and CEO** - **withinOrganization** - **Groupe Danone**
- Vice Chairman and COO** - **withinOrganization** - **Groupe Danone**

On the right side of the interface, there are two panels:

- Attributes of Groupe Danone:** Lists "Relations From: 5" and "Relations To: 3". It includes a scrollable list of attributes such as `hasMainAlias - Groupe Danone`, `stockExchangeIndex - DA`, `FISCAL_SALES - $12,897 mln.`, `FISCAL_NET_INCOME - $118 mln.`, `numberOfEmployees - 100,560`, and a `comment` describing the company as a global leader in cultured dairy products.
- History:** A list of nodes visited, including "Groupe Danone", "EVP, Finance", and "Emmanuel Faber".

At the bottom of the graph area, there are controls for "Zoom", "Rotate", and "Hyperbolic" views, along with a note: "Use horizontal scrollbar to zoom and rotate the graph".

Predefined Pattern Search

The screenshot shows a Microsoft Internet Explorer browser window titled "KIM WEB UI - Microsoft Internet Explorer". The address bar contains the URL "http://ontotest.sirma.bg/KIM/screen/PredefinedEntityPatterns.jsp". The page content is titled "Predefined Entity Patterns" and features a navigation menu on the left with items: Datastore, Entity Pattern Search, Predefined Patterns, Entity Lookup, Keyword Search, and About KIM. The main content area includes a dropdown menu for "Choose Predefined Entity Pattern" with the selected option "Person hasPosition Position withinOrganization Organization". Below this, a section titled "A Pattern about" contains three input fields: "Person" with the value "j*", "who has Position" with the value "spokesman", and "within Organization" with the value "IBM". A "Search for:" section has two buttons: "Documents" and "Entities". The footer of the page displays logos for "Powered by: GATE", "RDF SESAME", and "Lucene".

Pattern Search: Multiple-Entity Results

The screenshot shows a web browser window titled "KIM WEB UI - Microsoft Internet Explorer". The address bar contains the URL "http://ontotest.sirma.bg/KIM/screen/EntitiesResult.jsp". The main content area displays the "Entity Query Result" page. On the left, there is a navigation menu with items: "Datastore", "Entity Pattern Search", "Predefined Patterns", "Entity Lookup", "Keyword Search", and "About KIM". Below the menu, it says "Powered by:" followed by logos for "GATE", "RDF SESAME", and "Lucene".

The main content area features a table with the following data:

Person	Type	Tr	Position	Type	Tr	Organization	Type	Tr
<input type="checkbox"/> John Bukovinsky	Person		spokesman	Position		IBM	Company	+
<input type="checkbox"/> Joe Stunkard	Man		spokesman	Position		IBM	Company	+
<input type="checkbox"/> James Sciales	Man		spokesman	Position		IBM	Company	+
<input type="checkbox"/> Joseph Stunkard	Man		spokesman	Position		IBM	Company	+

Below the table, it indicates "1-4 of 4 Entities per page:" with a dropdown menu set to "30". There are three buttons: "Refine Query", "New Query", and "Edit Query". A "Hints:" section follows, containing two bullet points:

- » The **Type** column shows the most specific class the entity belongs to. For instance, the table could represent of class *Agent*. Then the entities in the table could be of number of different sub-classes of *Agent*, such as, *Persc*
- » The '+' sign in the **Tr** column indicates whether the entity was pre-populated in the knowledge base from a tr automatically extracted from some of the documents processed by the system. Due to natural limitations, the aut represent an incorrect or imprecise modelling of the world.

Pattern Search, Referring Documents

The screenshot shows a Microsoft Internet Explorer browser window titled "KIM WEB UI - Microsoft Internet Explorer". The address bar contains the URL "http://ontotest.sirma.bg/KIM/screen/DocumentsResult.jsp". The main content area displays a "Document Query Result" page. On the left, there is a navigation menu with the following items: Datastore, Entity Pattern Search, Predefined Patterns, Entity Lookup, Keyword Search, and About KIM. Below the menu, it says "Powered by:" followed by logos for GATE, RDF, SESAME, and Lucene. The main content area features a table with two columns: "Date" and "Title". The table contains six rows of data. Below the table, it indicates "1-6 of 6 Documents per page:" with a dropdown menu set to "30". There are also three buttons: "Refine Query", "New Query", and "Edit Query".

Date	Title
27/02/2003 21:22	IBM Cuts Small Percent of Software, Services Jobs
01/12/2003 15:14	IBM to Sell Software Specialized for Industries
05/01/2004 21:59	IBM Execs, S.Korea Officials Charged Over Bribery
07/11/2003 21:10	Report: South Korea Probes IBM Bribery Charge
07/03/2003 19:03	Applied Digital Sues IBM Over Microchip
11/02/2003 21:21	Dell: 1999 IBM Deal Falls Short of \$6 Billion Mark

Document Details

KIM WEB UI - Microsoft Internet Explorer


File Edit View Favorites Tools Help


Address <http://ontotest.sirma.bg/KIM/screen/DocumentDetail.jsp?UniqueID=524&selectedDoc=3> Go


KIM

- Datastore
- Entity Pattern Search
- Predefined Patterns
- Entity Lookup
- Keyword Search
- About KIM

Powered by:







Document Detail

Feature Name	Feature Value
TITLE	Report: South Korea Probes IBM Bribery Charge
SUBJECT	Technology
ORIGIN	NEW YORK
SOURCE	Reuters
UNIQUE_URL	http://www.reuters.com/newsArticle.jhtml?type=technologyNews&storyID=3780

Document Content

NEW YORK (Reuters) - South Korean prosecutors are investigating charges that International Business Machines Corp. (IBM) illegally bribed government officials, according to a story published on Thursday in the South Korean newspaper JoongAng Ilbo.

Armonk, **New York-based IBM** "is cooperating with the investigation," according to **IBM spokesman Jose** further on the nature of the investigation or the newspaper report.

According to a story posted on the JoongAng English language Web site, prosecutors raided three **IBM** offices in Armonk, N.Y., on Tuesday. A prosecutor told the newspaper that **IBM** gave cash to two of its business units to bribe government officials, the newspaper said.

The newspaper said one of the units was a joint venture with South Korea's LG called LG **IBM** that markets Winsol, markets its computer servers, the republic's natural gas requirements be met in full (18.5 billion cubic bcm in 2005, and 33 bcm per annum from 2010). This volume of natural gas should be supplied at the wholesale range in Russia.

Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - **Major Components, Architecture**
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- Wrap up

KIM Constituents

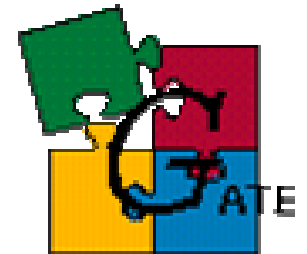
The KIM Platform includes:

- **KIM Server** – with a set of APIs for remote access and integration
- **Front-ends**, end-user facilities, ready to use:
 - Web UI – for zero installation access;
 - A light-weight semantic annotation plug-in for Internet Explorer.
- Massive Common **World Knowledge**
 - **Ontologies** (PROTON + KIMSO + KIMLO)
 - **KIM World KB**

KIM is based on ...

KIM is based on the following open-source platforms:

- **GATE** – the most popular NLP and IE platform in the world, developed at the University of Sheffield. Ontotext is its biggest co-developer.
www.gate.ac.uk and www.ontotext.com/gate



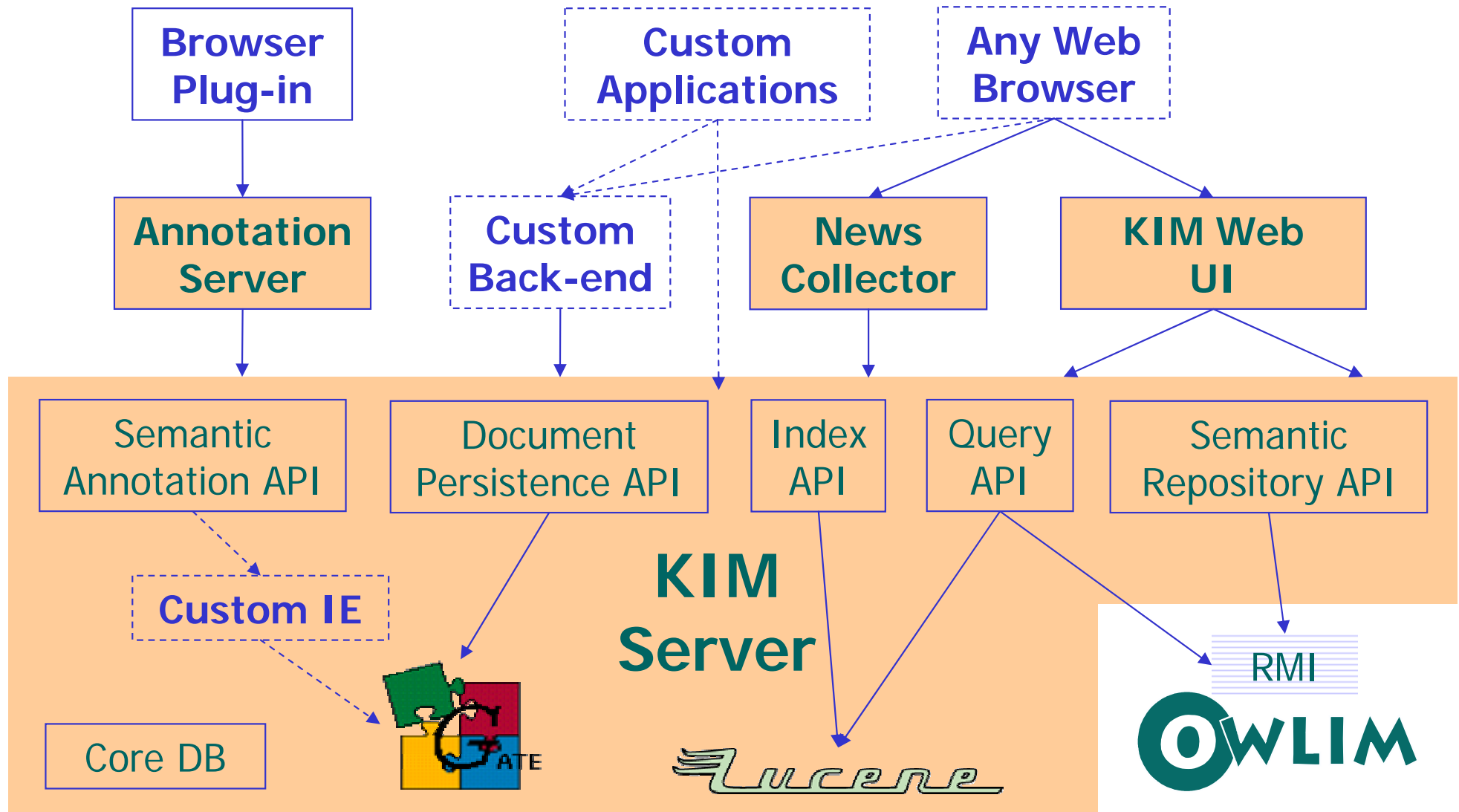
- **Sesame** – RDF(S) repository by Aduna B.V. Ontotext is its biggest co-developer.
www.openrdf.org



- **Lucene** – an open-source IR engine by Apache.
jakarta.apache.org/lucene/



KIM Architecture



Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - **Information Extraction: People Search For People**
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- Wrap up

People Search for People

A recent large-scale human interaction study on a personal content IR system, carried out by Microsoft ([10]), demonstrated that:

“The most common query types in our logs were People/places/things, Computers/internet and Health/science. In the People/places thing category, names were especially prevalent. Their importance is highlighted by the fact that **25% of the queries involved people’s names** In contrast, general informational queries are less prevalent.”

[10] Dumais S., Cutrell E., Cadiz J., Jancke G., Sarin R. and Robbins D. *Stuff I've Seen: A system for personal information retrieval and re-use*. In proc. of SIGIR'03, July 28 – August 1, 2003, Toronto, Canada, ACM Press, pp. 72-79.

Semantic Metadata in KIM

- Provides a specific **metadata schema**,
 - focusing on named **entities (particulars)**,
 - also number and time-expressions, addresses, terms, etc.
 - everything “specific”, apart from the general concepts.
- Defines specific **tasks for generation and usage** of metadata,
 - which are well-understood and measurable.
- Why not metadata about general things (universals)?
 - Even partial descriptions are too complex (think of Cyc and WordNet)
 - But one can easily extend KIM in this direction
- The particulars seem to provide a good **80/20 compromise**
 - They also appear to be key “characteristic features” of texts

Semantic Annotation of NEs

A Semantic Annotation of the named entities (NEs) in a text includes:

- **recognition of the type** of the entities in the text
 - out of a **rich taxonomy of classes** (not a flat set of 10 types);
- **identification** of the entities, (identity resolution):
 - this problem is similar to “record linking”, “co-reference resolution”

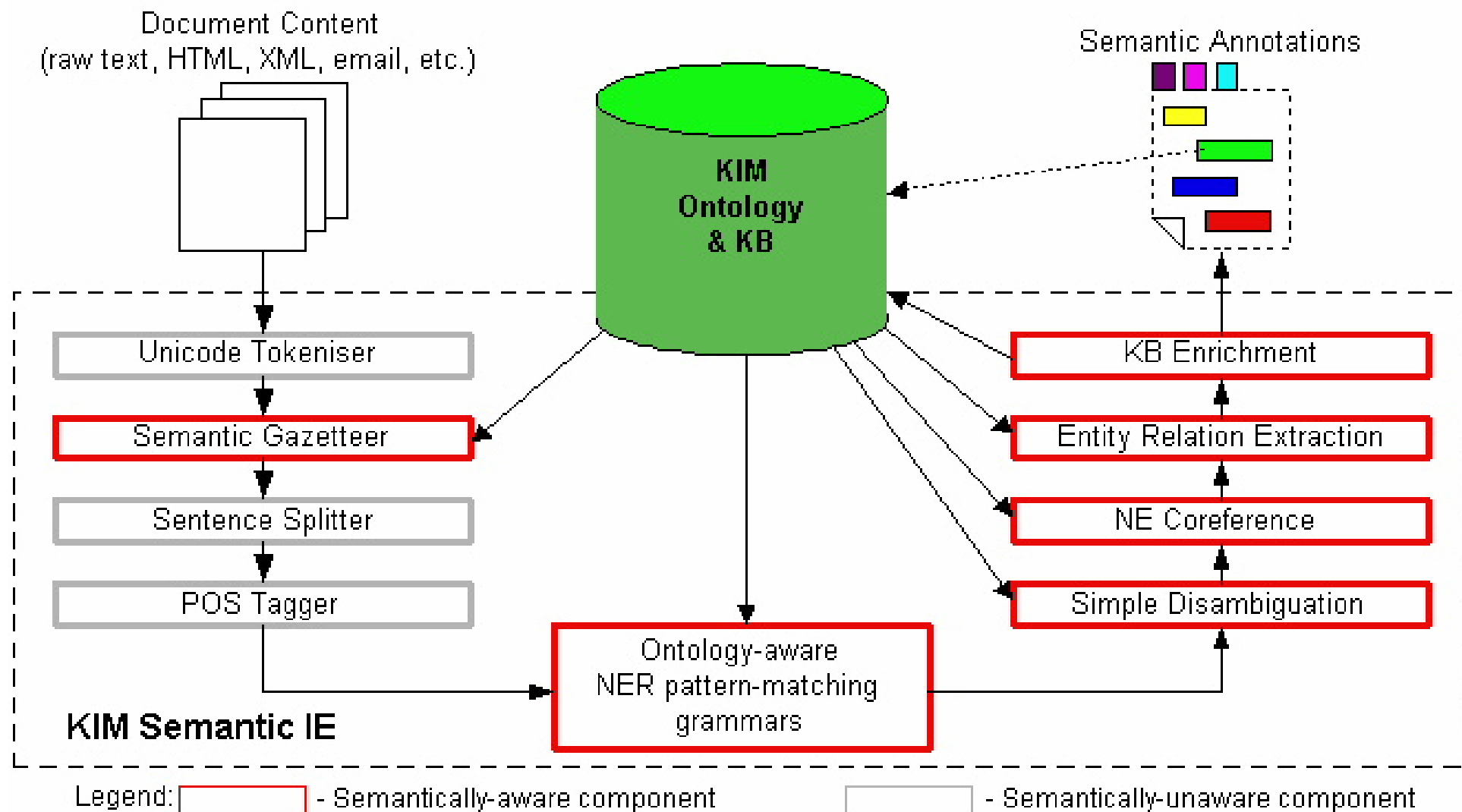
The **traditional (IE-style) NE recognition** approach results in:

```
<Location>Barbados</Location>
```

The **Semantic Annotation of NEs** results in:

```
<Island ID="http://...#Island.1234">  
    Barbados  
</Island>
```

KIM Information Extraction Pipeline



Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - **Massive “World Knowledge” in the Background**
 - Scalability, KIM’s Cluster Architecture
- Wrap up

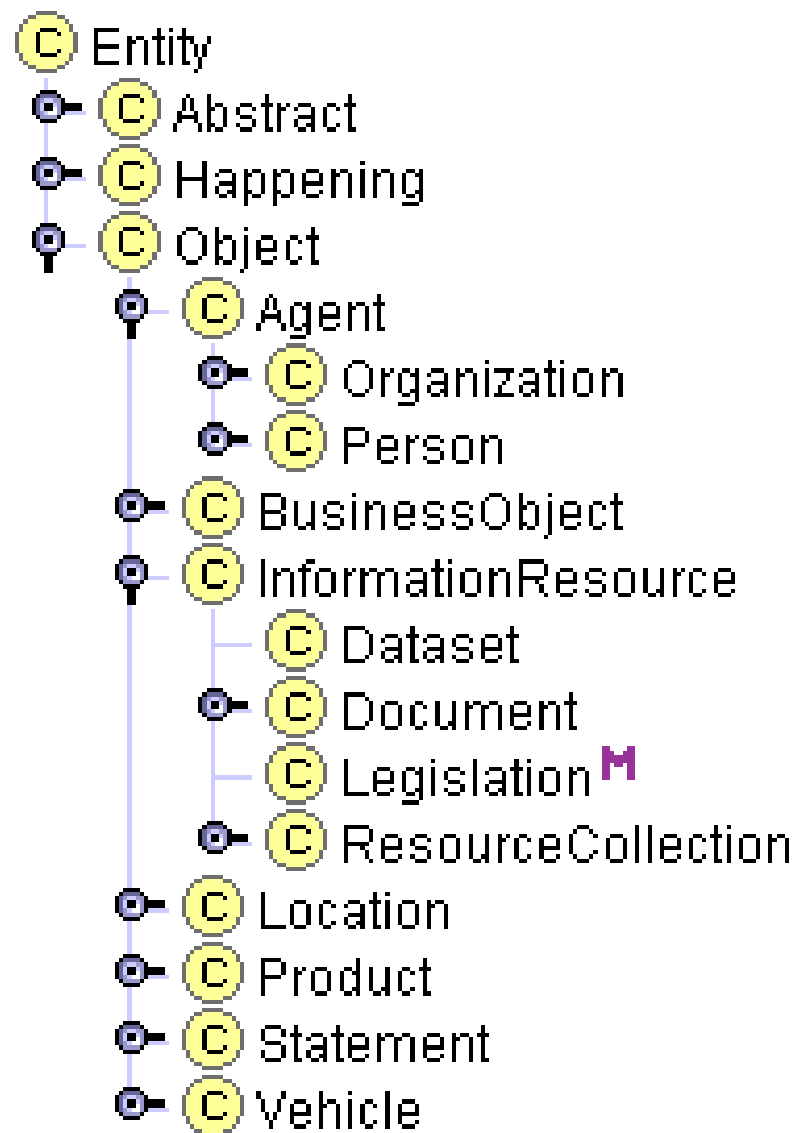
World Knowledge in KIM

Rationale:

- provide common knowledge about **world entities**;
- **KIM bets on scale** and avoids heavy semantics;
- minimum modeling of common-sense, almost no axioms;
- Represented in OWL Lite (actually, OWL DLP – a tractable dialect)

Ontologies

- **PROTON** - a light-weight upper-level ontology;
- **250 NE classes**;
- **100 relations and attributes**;
- covers mostly **NE classes**, and to a smaller degree general concepts;
- **Modules**: System, Top (on the right), Upper, KM
- Couple of KIM specific ontologies: KIMSO, KIMLO
- A common basis for domain extensions



<http://proton.semanticweb.org/>

Ontologies II

KIM WEB UI - Microsoft Internet Explorer

Address <http://ontotest.sirma.bg/KIM/screen/ontobrowse.jsp>

KIM **Ontology**

Available Entity Classes

- Entity
 - Abstract
 - Happening
 - Event
 - Situation
 - TimeInterval
 - Object
 - Agent
 - Organization
 - Person
 - BusinessObject
 - InformationResource
 - Location
 - Statement
 - Vehicle

Organization, a Class

Property	Value
comment	"Organization is a group, which is established such that certain known relationships and obligations exist between the members, and/or between the organization and its members, and/or between the organization and 'outsiders' (individuals or groups). Organization includes both informal and legally constituted organizations. Organizations can act as agents -- to undertake projects, enter into agreements, own property, etc. Most organizations have names. Almost all have at least two members."
subClassOf	Agent

Powered by:

Copyright © 2004 Ontotext Lab, Sirma AI, Bulgaria

KIM World KB

A **quasi-exhaustive** coverage of the **most popular entities** in the world ...

- What a person is expected to have heard about that is beyond the horizons of his country, profession, and hobbies.
- Entities of **general importance** ... like the ones that appear in the **news** ...

KIM “knows”:

- **Locations**: mountains, cities, roads, etc.
- **Organizations**, all important sorts of: business, international, political, government, sport, academic...
- Specific **people**, etc.

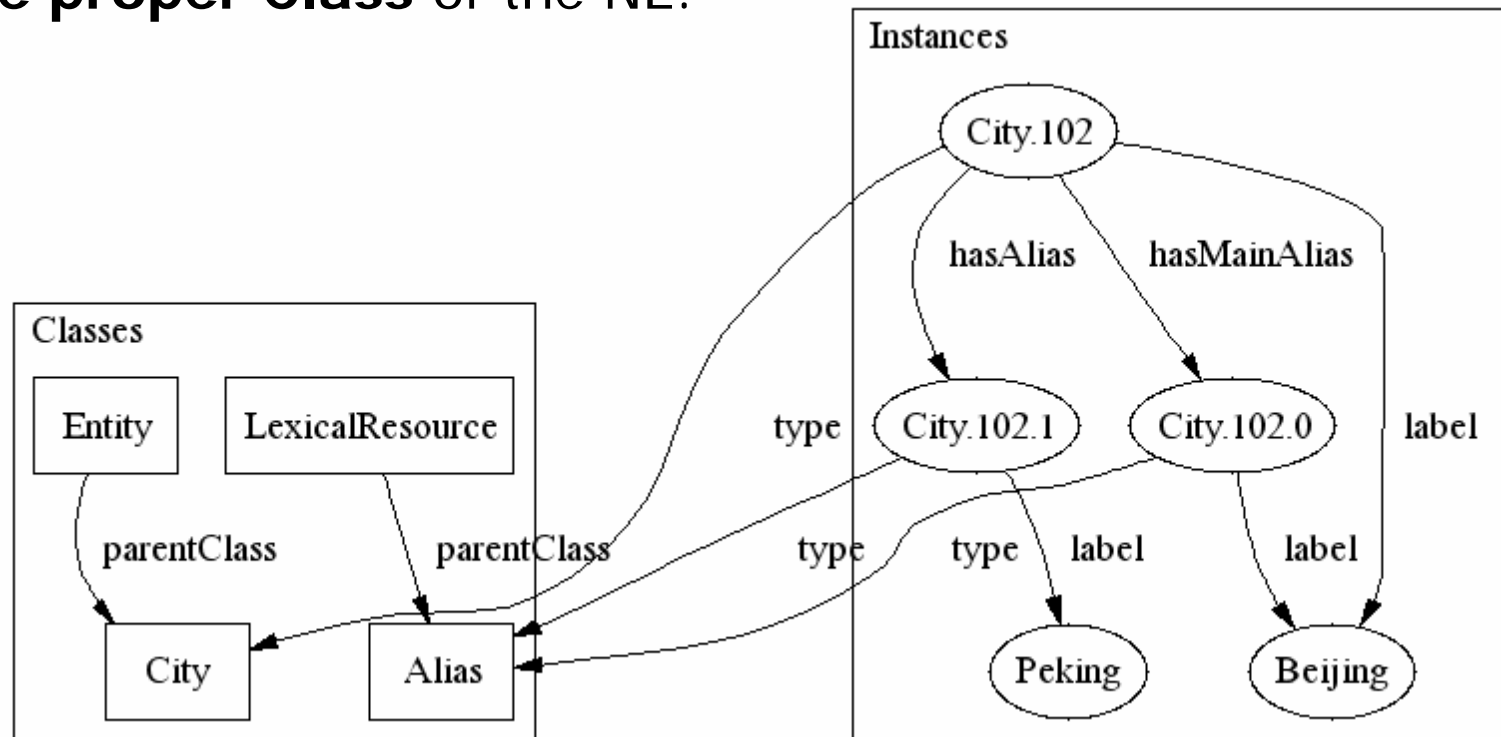
KIM World KB: Content

- **Collected** from various sources, like **geographical** and **business intelligence** gazetteers.
- So, it is all predefined? ... **NO, KIM learns** from the texts.
- The **KIM World KB only provides the seed**, the “common culture”, which is:
 - basic, being referred to often, so it has to be modeled well;
 - hard to extract from regular texts, because the authors expect the readers to know it:
 - in reports and news articles, nobody bothers to explain what “Asia” or “United Nations” stands for.

KIM World KB: Entity Description

The NE-s are represented with their **Semantic Descriptions** via:

- **Aliases** (*Florida & FL*);
- **Relations** with other entities (*Person hasPosition Position*);
- **Attributes** (*latitude & longitude* of geographic entities);
- **the proper Class** of the NE.



The Scale of KIM World KB

RDF Statements	Small KB	Full KB
- explicit	444,086	2,248,576
- after inference	1,014,409	5,200,017
Instances		
- Entity:	40,804	205,287
- Location:	12,528	35,590
- Country:	261	261
- Province:	4,262	4,262
- City:	4,400	4,417
- Organization:	8,339	146,969
- Company:	7,848	146,262
- Person:	6,022	6,354
- Alias:	64,589	429,035

Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - **Scalability, KIM’s Cluster Architecture**
- Wrap up

KIM Scaling on Data

- To manage ontologies and KBs, KIM uses **OWLIM**
 - OWLIM is a high-performance Sesame SAIL with OWL inference
 - **SwiftOWLIM is the fastest OWL machine**, even on desktop PC
 - It can load and infer over 7M statements, LUBM(50), in 6 min.
 - Processing speed 40,000 Statement/sec.
 - **BigOWLIM is the most scalable OWL machine**
 - It can load and infer over 1 Billion st., LUBM(8000), in 69h!
- On average, each entity is described by 10 RDF statements
 - I.e. BigOWLIM can handle 100 million entities;
- KIM can index and manage 1M documents on \$5000-worth server

KIM Cluster Architecture

- Scalability has been identified as a critical issue for:
 - the processing of large volumes of data, so that **statistical information extraction** (IE) methods could be designed and trained;
 - the enabling of **public metadata-on-demand services**
- Extensive scaling should be enabled, and there comes the KIM Cluster Architecture. Here are some of its features:
 - support for a virtually unlimited number of annotators (the components, performing the computationally most expensive processing);
 - **centralized ontology storage** and querying;
 - centralized meta-data (annotations) and document storage, indexing, and querying;
 - **support for multiple crawlers** (or other data sources);
 - **dynamic reconfiguration** of the cluster (e.g. starting new crawlers or annotators on demand).

KIM Cluster Architecture III

CLUSTER CONSOLE

Table View
Graph View
KIM Web Interface
Statistics
Refresh
About SWAN

Cluster Control

Name: **master-kim**
» Running on: ontotest
» Status: **running**
» Type: master-kim
» Dependencies: configuration: kim-cluster
» Properties: document stored: 0

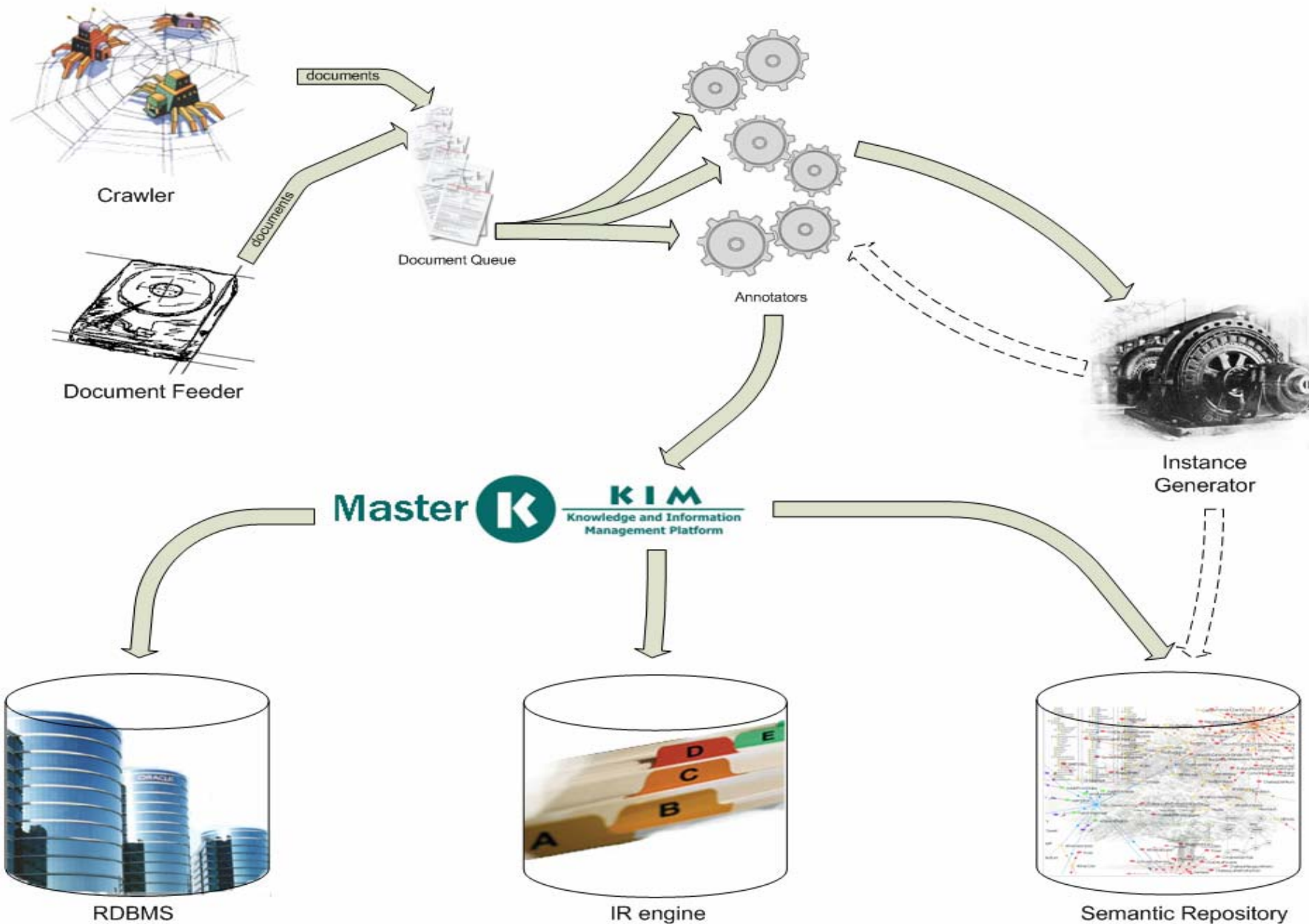
Powered by:

KIM
Knowledge and Information Management Platform

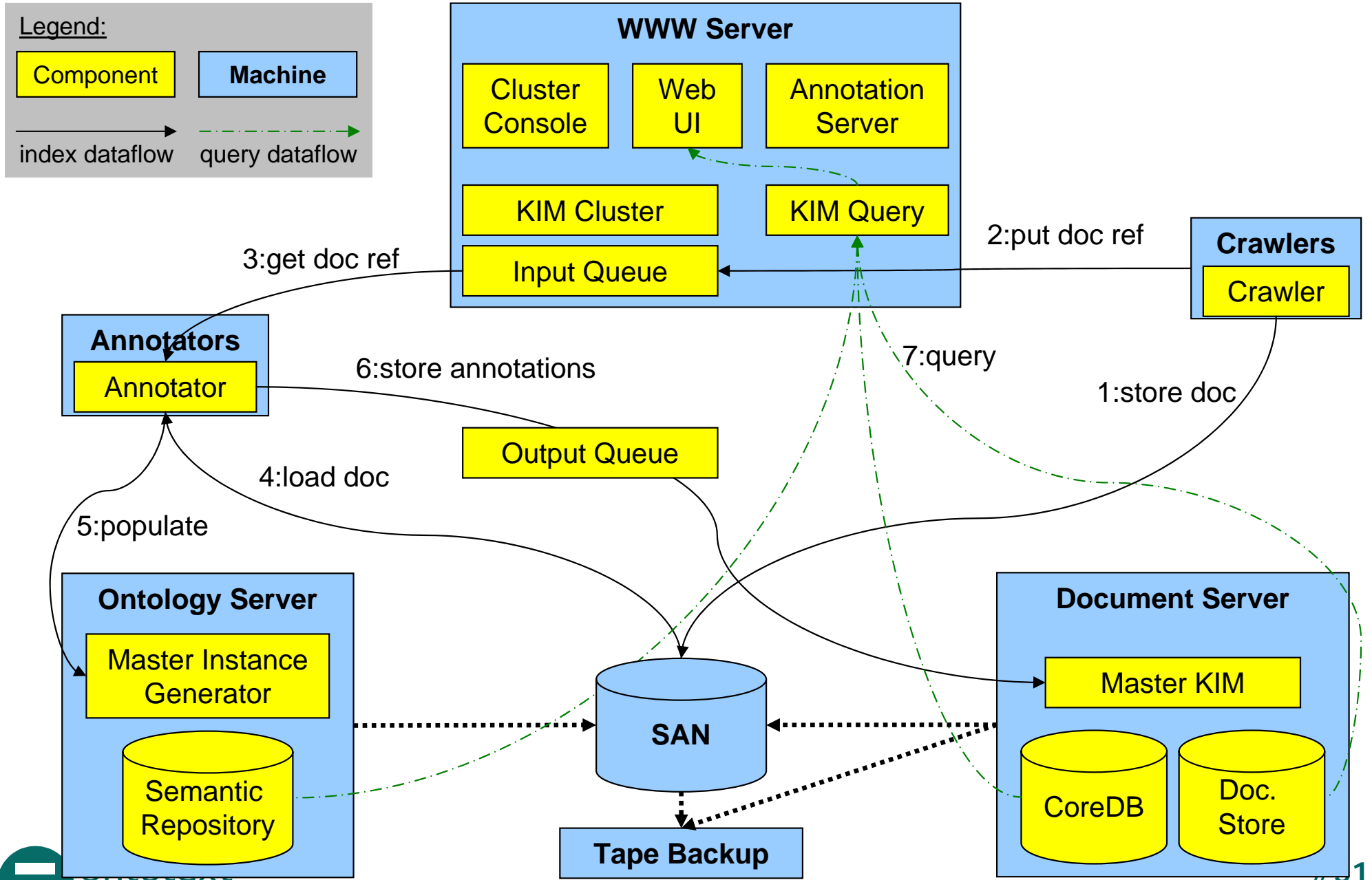
ATE

Name	Status	Used Memory	Free Memory	Additional Properties
localhost	available	754 MB	1230 MB	No additional properties
semantic-repository	available			
rosem	unavailable	0 MB	0 MB	No additional properties
annotator/rosem	unavailable			
ontotest	available	72609 MB	42591 MB	Freq, GHz: 2.4 RAM, MB: 2048 CPU count: 1
master-kim	running			
peter	unavailable	846 MB	1138 MB	No additional properties
annotator/peci	unavailable			
192.168.128.219	available	13430 MB	9874 MB	No additional properties
document-queue	available			
master-ig	running			

Cluster Architecture – An Overview



Sample Cluster Configuration



Presentation Outline

- What: Functionality
 - Text-Mining, Semantic Annotation, and Hyper-linking
 - Co-occurrence and Popularity Timelines
 - Combining FTS, Structured Queries, and Inference
- How: Architecture & Implementation
 - Major Components, Architecture
 - Information Extraction: People Search For People
 - Massive “World Knowledge” in the Background
 - Scalability, KIM’s Cluster Architecture
- **Wrap up**

General-Purpose and Robust

KIM is:

- **open-domain** – take an **arbitrary document** and annotate it;
- **robust** – it processes **thousands of documents** every day:

the **News Collector** uses KIM to **annotate** and **index** the **news** that are **daily emitted** by a **dozen** of the **leading news wires**

- intended to be used as a back-end infrastructure:
 - like the DBs and the Indexing engines;

Applications, which are built on KIM, take its “basic intelligence” and “educate” it for the particular task, domain, context...

- e.g., a company would probably extend the KB with data from its CRM system.

KIM Applications & Customization

KIM can be customized by:

- changing or **extending the ontology**;
- adding more world or **domain knowledge**;
- developing new **GATE-based IE applications**;
- tuning the **lexical resources**;
- implementing new front-end tools.

Wrap Up

KIM is a **platform** for:

- **semantic annotation,**
- **ontology population,**
- **semantic indexing** and **retrieval,**
- providing an **API** for **remote access** and **integration,**
- based on **Information Extraction (IE)** using **mature HLT (GATE).**

KIM offers:

- **text-mining** powered by **massive world knowledge;**
- **robust, scalable, general-purpose,** off-the-shelf platform!

Thank You

Give KIM a try

<http://www.ontotext.com/kim>

Download the Internet Explorer annotation plug-in

Play with the Public annotation and search services