



Ontotext

Knowledge and Language
Engineering Lab of Sirma



Towards Semantic Web Information Extraction

...or the snowball to cause the avalanche

The web is huge...(cliché)

...in order to become “**Semantic**” it will need **masses of metadata** associated with its resources...

- Most of this metadata will be **automatically acquired**.
- It should be **machine-readable** (wrt some formal knowledge)

Named Entities (NE) are considered:

people, organizations, locations, and others referred by name.

May also include scalars and expressions:

numbers, amounts of money, dates, etc. (NUMEX, TIMEX)

Our **hypothesis** is that the **named entities** (and the relations between them) mentioned in a resource constitute **an important part of its semantics**.

Semantic Annotation of the NEs in a text includes:

- **Recognition of the type** of the entities in the text
- **Identification of the entity individual**

...the **traditional NER** approach results in:

<Person>Lama Ole Nydahl</Person>

...the **Semantic Annotation of NEs** should result in something like the following:

*<ReligiousPerson ID="<http://..kim/Person111111>">
Lama Ole Nydahl</ReligiousPerson>*

The **K**nowledge and **I**nformation **M**anagement Platform is a **software product, platform** or **SDK** that provides:

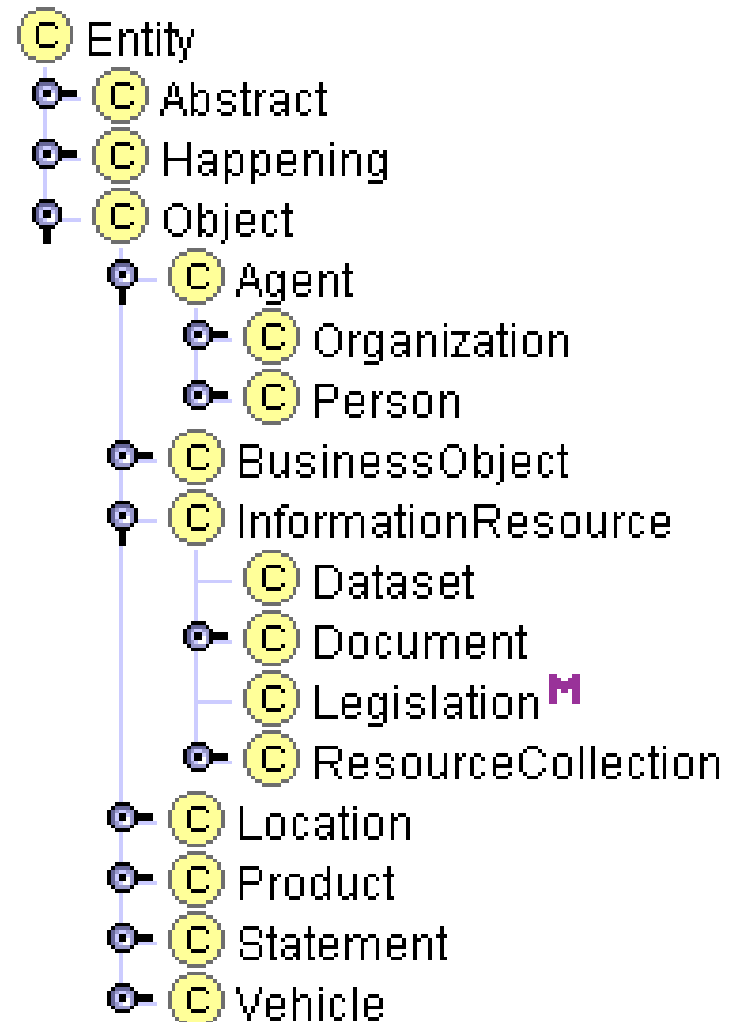
- **Automatic Semantic Annotation** of NEs (and relations between them)
- **Ontology Population** with NE individuals and relations
- **Indexing** and **Retrieval** wrt NEs
- **Query** and **Navigation** over the **Formal Knowledge**

The KIM Platform includes:

- **KIM Ontology (KIMO)**
- **KIM World KB**
- **KIM Server** – with API for remote access and integration
- **Front-ends:** KIM Web UI, Plug-in for Internet Explorer, and KB Explorer

- **light-weight upper-level ontology**
- **250 NE classes**
- **100 relations and attributes:**
- covers mostly **NE classes**, and ignores general concepts
- includes classes representing **lexical resources**

www.ontotext.com/KIM/kimo.rdfs



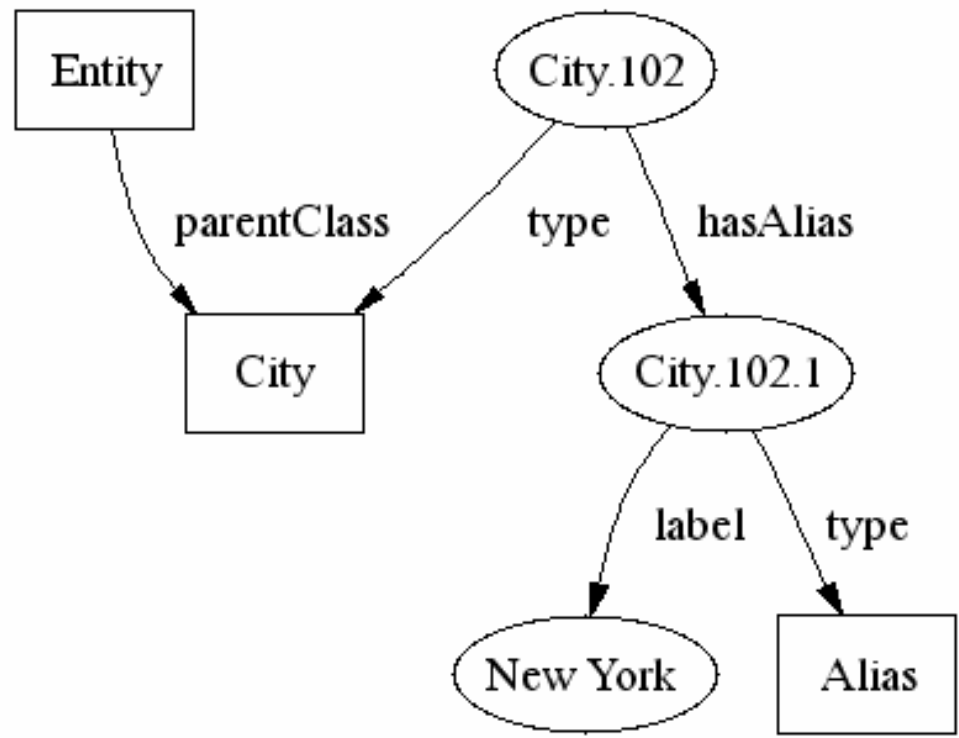
...**quasi-exhaustive** coverage of the **most popular entities** in the world...

...entities of **general importance**...like the ones that appear in the **news**...

At present KIM KB consists of about **200 000 entities**:
50 000 locations, 130000 organizations, 6000 people, etc.

The NEs are represented in KIM World KB with their **Semantic Descriptions** consisting of...

- **Aliases** (*Florida & FL*)
- **Relations** with other entities (*Person hasPosition Position*)
- **Attributes** (*latitude & longitude* of geographic entities)
- **Proper class** of the NE

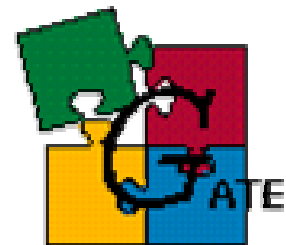


APIs for:

- **Semantic Annotation**
- **Document Persistence**
- **Indexing & Retrieval** of documents wrt NEs
- **Semantic Repository Access & Exploration**

KIM is based on the following open-source platforms:

- **GATE** – leading NLP and IE platform developed in the University of Sheffield.



Ontotext contributed to releases 2.0 and 2.1.

www.gate.ac.uk and www.ontotext.com/gate

- **Sesame** – RDF(S) repository by Aidadministrator b.v. Ontology Middleware and Custom Inference



by Ontotext as extensions of Sesame.

sesame.aidadministrator.nl, www.ontotext.com/omm

- **Lucene** – open-source IR engine from Apache.

jakarta.apache.org/lucene/

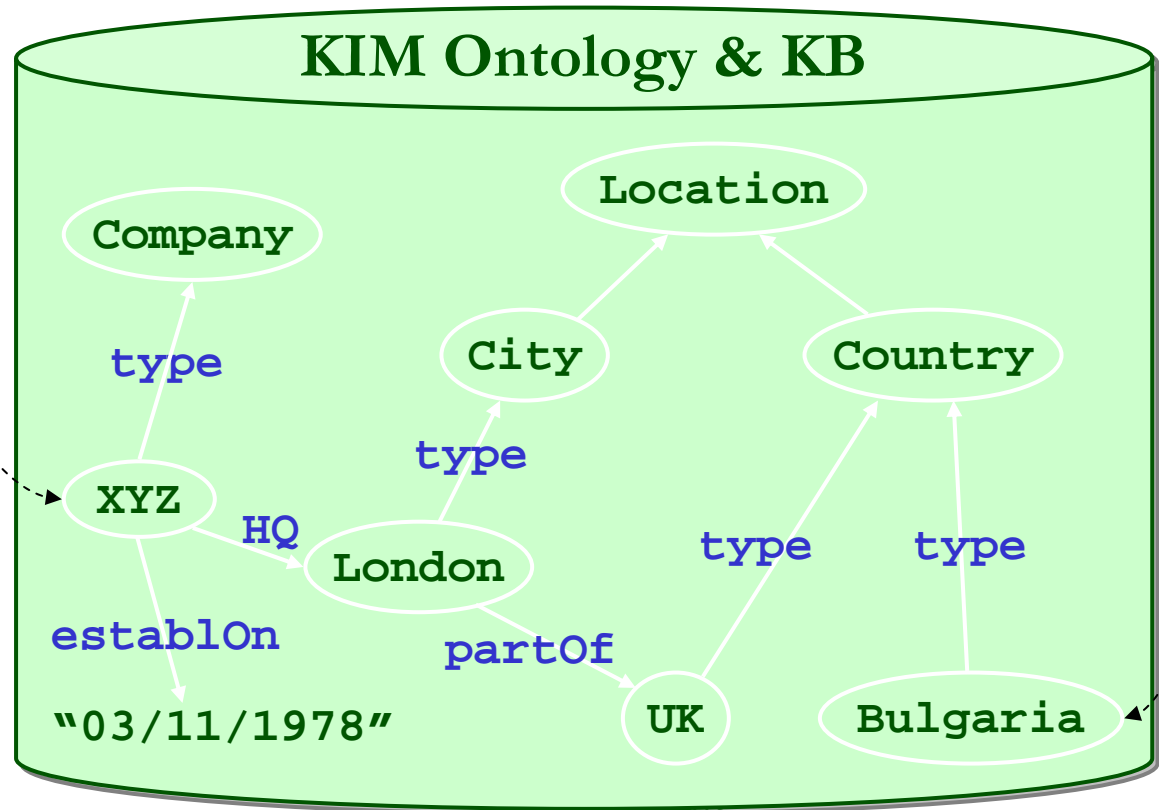


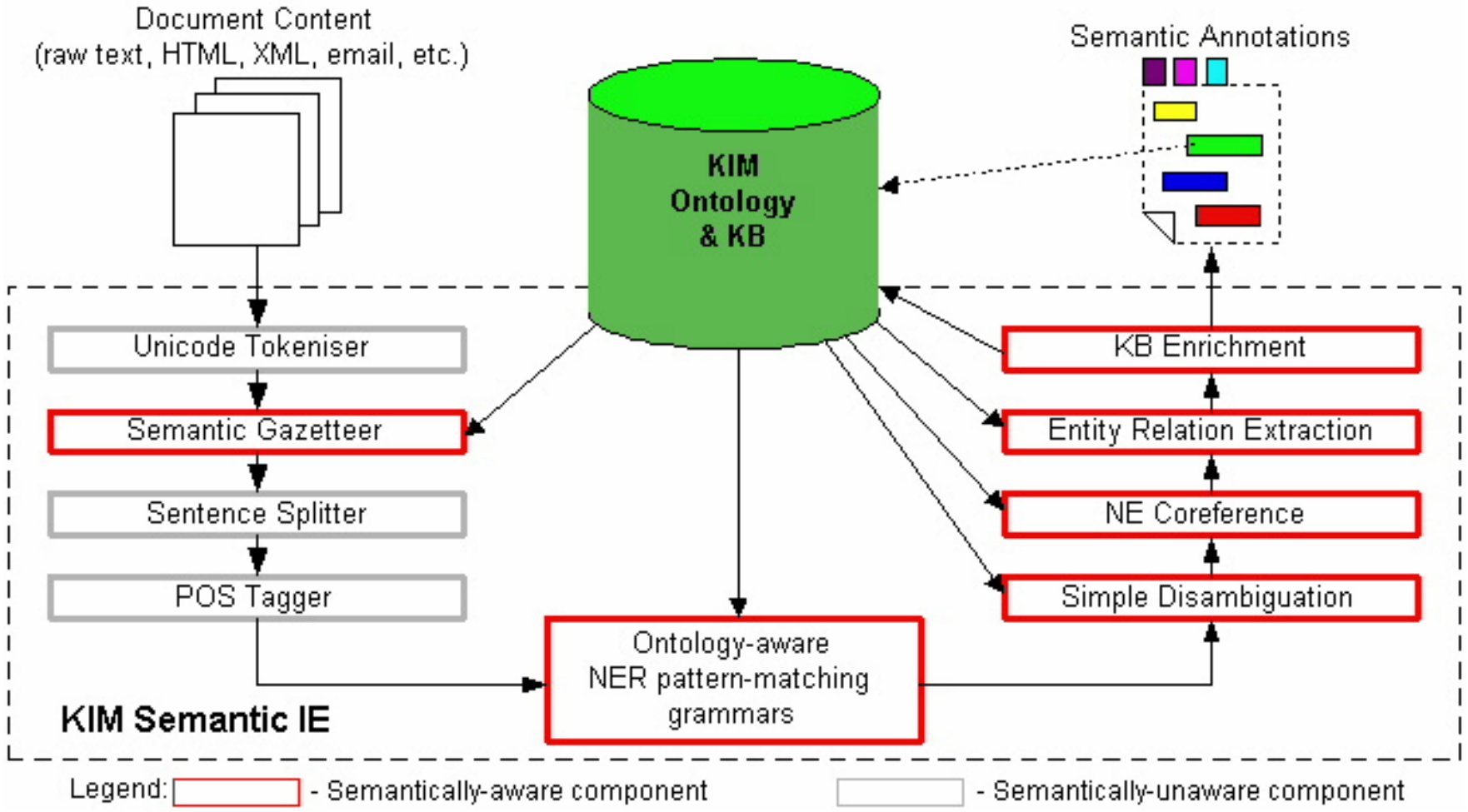
What's so special ?

- Based on **GATE** – a good example of adapting existing HLT's for the Semantic Web.
- **IE** based on **massive world knowledge**
- **Cheap** acquisition of **masses of metadata**
- Recognition and **Identification** of the NEs
- **IE** supported by a **Semantic Repository**:
 - Containing **lexical** and **gazetteer resources**
 - **Annotations** referring to **Entity Descriptions**
- **Ontology Population** with the **newly recognized entities & relations**

Semantic Annotation Diagram

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





Evaluated over 3 human-annotated corpora of **news** articles:

International Business News, International Political News, and UK Political News (~500 articles):

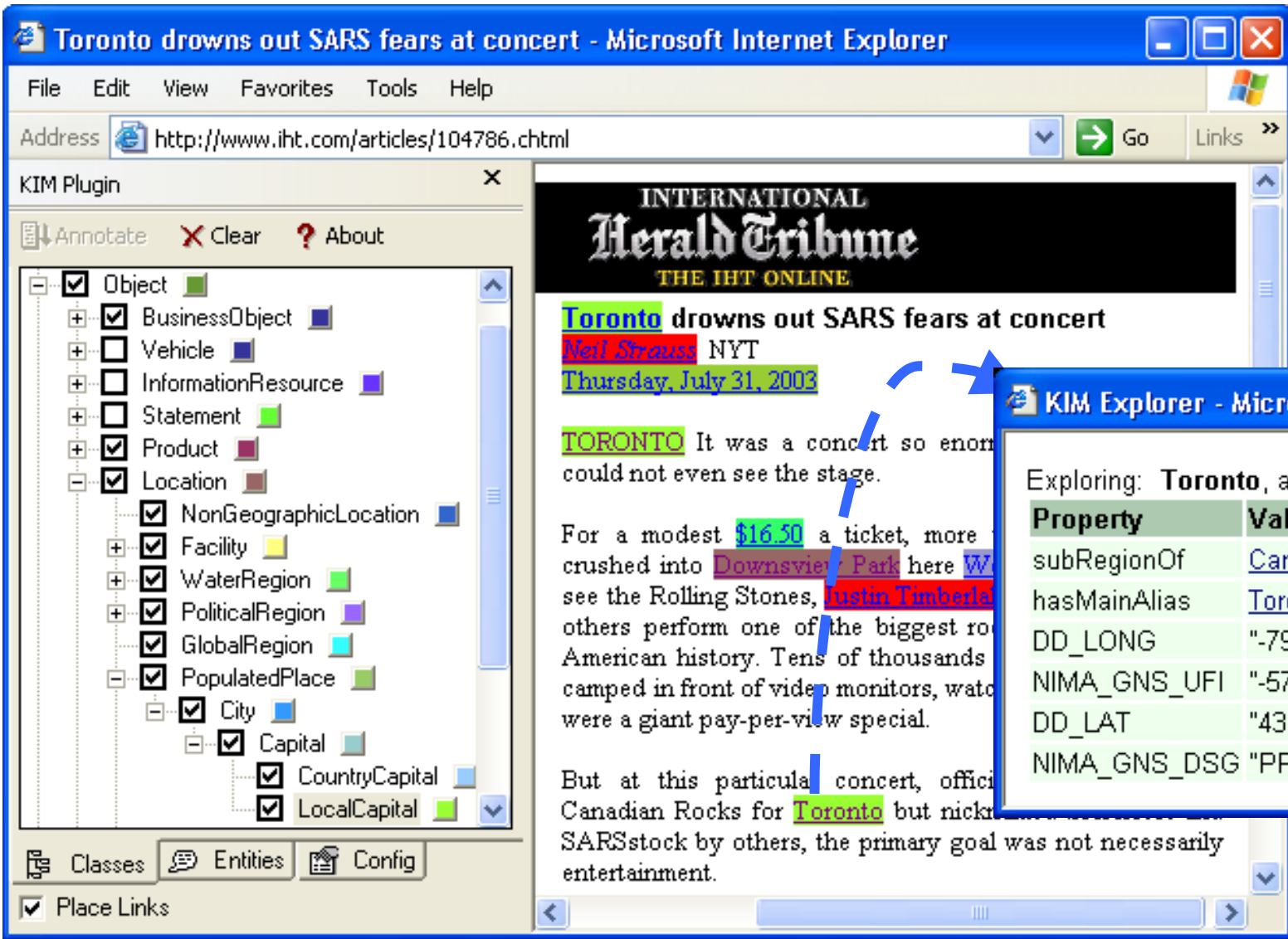
Precision 90%, Recall 83% wrt the standard NE types

...but these metrics are **not representative** for semantic annotation...

There are no established metrics for semantic annotation:

- No human-annotated corpora with precise class and instance information
- No metrics for various partial matches
 - When a **more specific class** is recognized
 - When a **more general class** is recognized
 - When the class is correctly recognized, but the individual **entity is not correctly identified.**

- **Smooth traversal** between the **annotated content** and the **associated formal knowledge**
- **Semantically enhanced Information Retrieval**
- More complex **knowledge acquisition** based on the identification of **relations between entities** and **entity attributes**



Toronto drowns out SARS fears at concert - Microsoft Internet Explorer

Address: <http://www.iht.com/articles/104786.html>

KIM Plugin

Annotations: Annotate, Clear, About

- Object
 - BusinessObject
 - Vehicle
 - InformationResource
 - Statement
 - Product
 - Location
 - NonGeographicLocation
 - Facility
 - WaterRegion
 - PoliticalRegion
 - GlobalRegion
 - PopulatedPlace
 - City
 - Capital
 - CountryCapital
 - LocalCapital

KIM Explorer - Microsoft Internet Explorer

Exploring: Toronto, a LocalCapital

Property	Value
subRegionOf	Canada
hasMainAlias	Toronto
DD_LONG	"-79.4166667"
NIMA_GNS_UFI	"-574890"
DD_LAT	"43.6666667"
NIMA_GNS_DSG	"PPLA"

INTERNATIONAL Herald Tribune THE IHT ONLINE

Toronto drowns out SARS fears at concert
 Neil Strauss NYT
 Thursday, July 31, 2003

TORONTO It was a concert so enormous that some people could not even see the stage.

For a modest \$16.50 a ticket, more than 100,000 fans crushed into Downsview Park here Wednesday night to see the Rolling Stones, Justin Timberlake and others perform one of the biggest rock concerts in American history. Tens of thousands of fans camped in front of video monitors, watching the concert on a giant pay-per-view special.

But at this particular concert, officials were concerned about SARS stock by others, the primary goal was not necessarily entertainment.

The standard IR need is:

“give me the resources that contain the words/ stems (f.e. company, telecommunication)...”

KIM provides **indexing & retrieval wrt NEs**

This enables the specification and satisfaction of semantically enhanced information needs.

KIM allows us to specify the NEs we’re interested in, and restrict them by their attributes and relations to other entities. So we could ask:

“Give me all documents that mention a company from the telecommunications industry sector...”

KIM IE might be customized by:

- Extending/changing the ontology
- Adding more domain/world knowledge
- Modification of the lexical resources
- Developing specific IE applications based on GATE and then plugging them in KIM



Ontotext

Knowledge and Language
Engineering Lab of Sirma

Thank You



Give **KIM** a try:

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