



Ontotext

Knowledge and Language
Engineering Lab of Sirma



Semantic, Annotation, Indexing and Retrieval

*or One Way to Define and Satisfy
Information Needs in the Semantic Web*

11/6/2003

- Introduction
 - Information access methods
 - Semantic Web and Metadata
 - Our Approach
 - Information Extraction
 - Semantic Annotation
 - Indexing & Retrieval
 - Model and Representation
 - KIM Platform: Implementing the Vision
-



Introduction



- Semantic Web is about **adding formal semantics** to the web for the purpose of more efficient access and management.
 - Its vitality depends on the presence of **critical mass of metadata**. The acquisition of this metadata is a major challenge.
 - Our vision is that **fully automatic methods** for semantic annotation should be researched and developed.
 - The related **design and modeling questions** should be faced and resolved
 - The **enabling resources and infrastructure** to be provided
-

- An *information access method* is a paradigm providing:
 1. definition of the information need (the *question*) and
 2. the means to satisfy it (the nature of the *answer/result*)
- IR has a basic information access method, **Document Retrieval**:
 - Need: expressed as a set of tokens of interest and
 - Satisfaction: list of documents relevant to those tokens
- Same with the relational databases:
 - Need: defined as an SQL query and
 - Satisfaction: result table.

What makes a Good Information Access Method



It should combine:

- **conscious user needs**
 - At least needs which can be understood and adapted to replace others, harder to satisfy.
 - **sound scientific theory**
 - **robust technology** which can implement applications based on the theory and efficient enough in satisfying the needs.
 - At the end of the day, this is a question of **efficiency and expectations management**.
 - keyword-based search engines are far not perfect: the information need is poorly defined (one would prefer Q/A) and imprecisely satisfied.
-

What makes a Good Information Access Method



The search engines are popular because they meet number of conditions critical for wide acceptance of an information access method in web context:

- Significantly **improve the efficiency** of accessing the content;
- **Do not require additional skills, effort**, discipline, good will, and correctness from the authors;
- To offer somehow **predictable behavior** and performance.

- Semantic Web has **no information access methods**
 - It is nice to have metadata, now what?
- **How is the information need defined and satisfied within the Semantic Web?**

The previous question is badly stated, as the current web :

- the Semantic Web may not have a single access method, thus single approach for definition of the information need.

More relevant questions are:

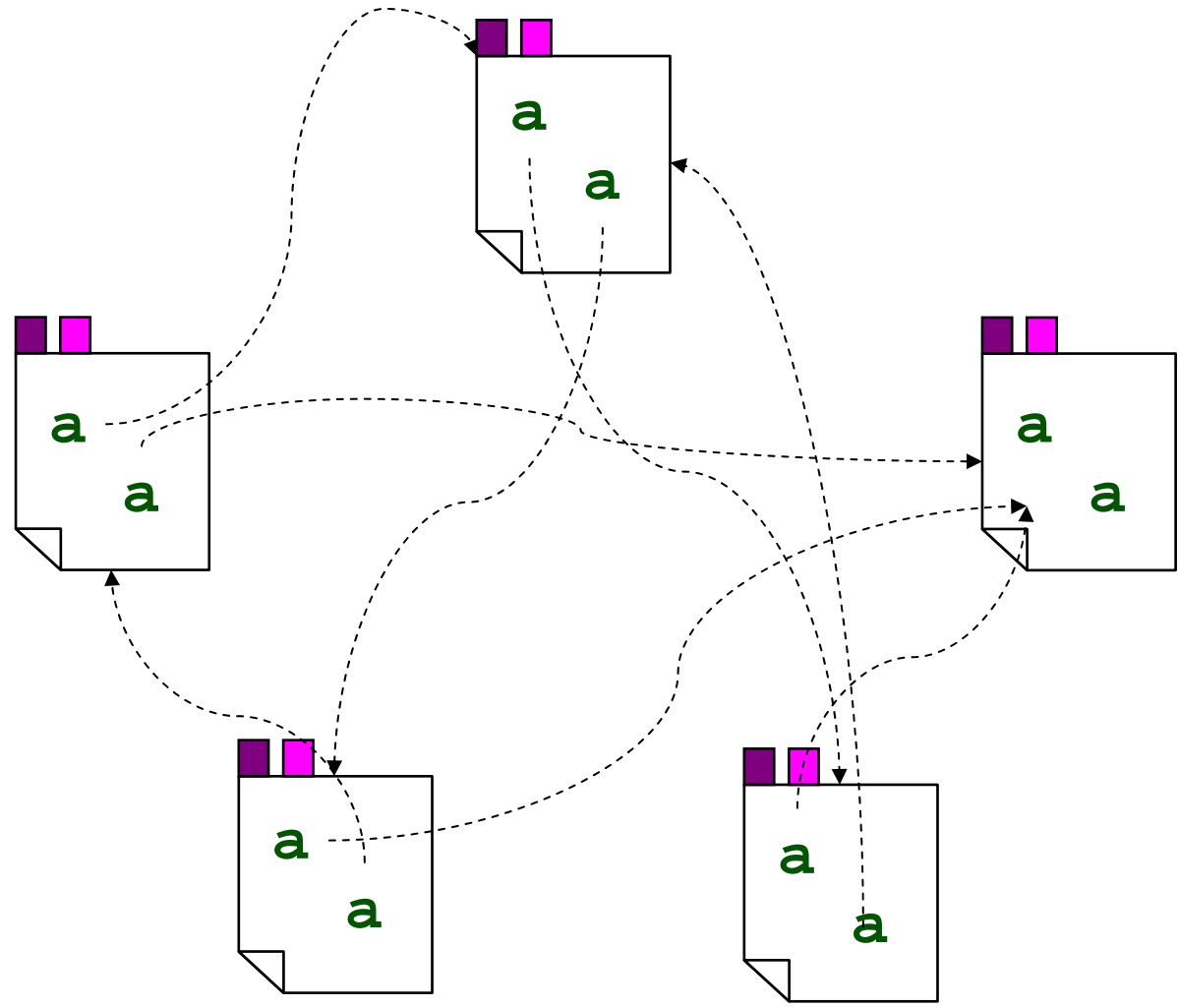
- How does the Semantic web extend the existing access methods?
- What new access methods become feasible?

- Introduction
 - Information access methods
 - **Semantic Web and Metadata**
 - Our Approach
 - Information Extraction
 - Semantic Annotation
 - Indexing & Retrieval
 - Model and Representation
 - KIM Platform: Implementing the Vision
-

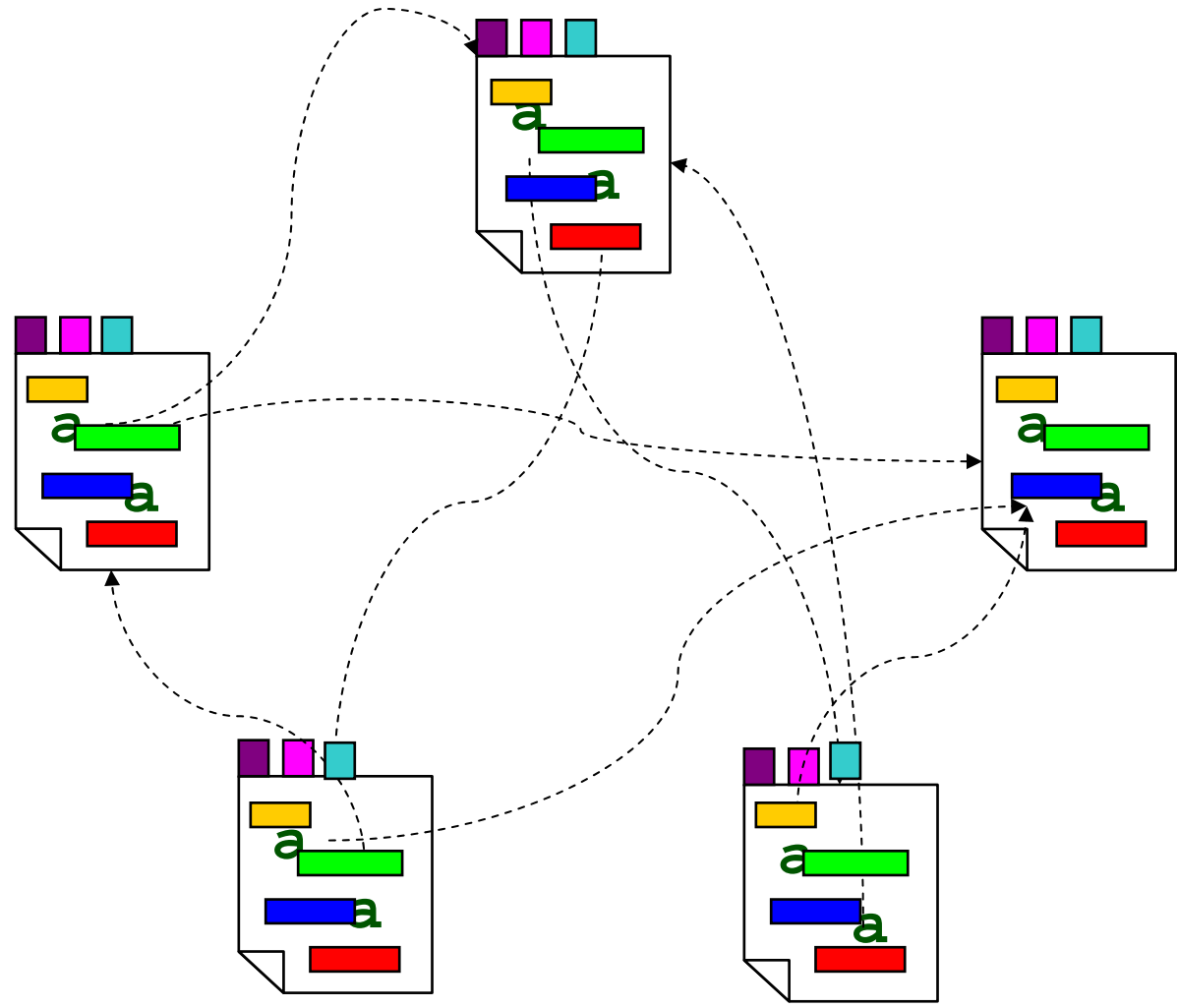
- **Semantic Web** at present offers
 - **A high-level vision**
 - **Low-level standards**
 - It is like:
 - Knowing you want to entertain and it probably requires money
 - But without the idea for Beach, Surf, Hiking, Theater, Bar
 - And, of course, without specific beaches, surfs, theaters, and bars
 - Having some ways to make and manage money, but no specific ideas and opportunities on how to spend them
-

- Saying:
 - There should be metadata! and
 - It should be in RDF(S) and related standards
 - Does not provide enough guidance for development and usage
 - **What is missing?**
 - Well-defined, measurable, **widely understood tasks**
 - Specific practices and approaches for performing them
 - Tools, resources, industry support
 - But the **tools come after the tasks and the problems**
-

The Current Web



The Semantic Web?



What does semantic mean?



Answer 1: colourful

Answer 2: nice

Answer 3: formal

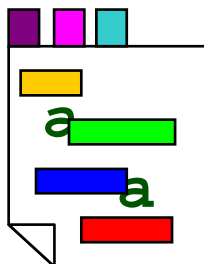
Answer 4: expensive

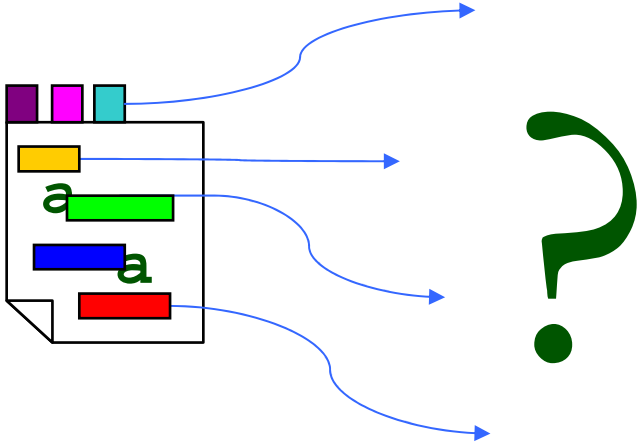
Answer 5: symbolic

Answer 6: efficient

Answer 7: metadata

Answer 7: anti-terroristic





OK, should have something to do with **METADATA**.

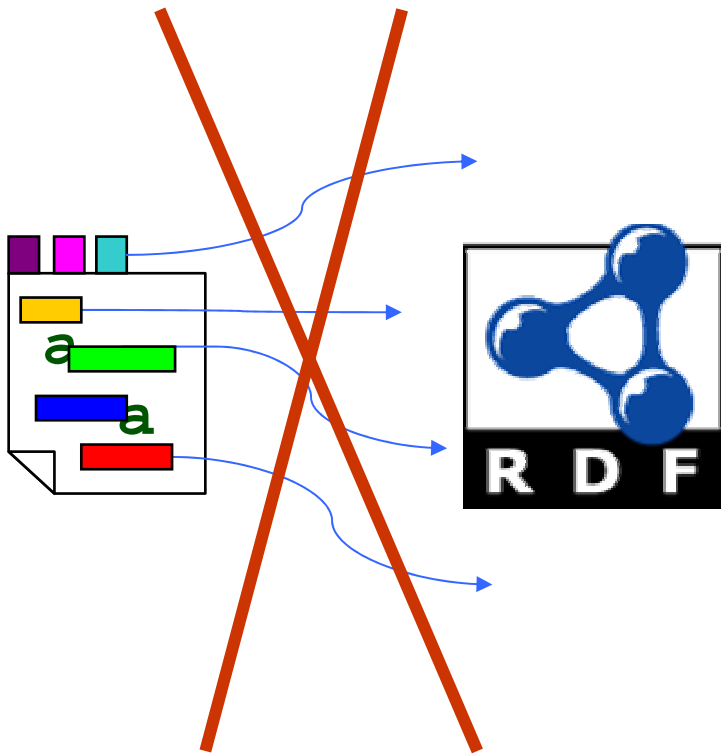
Any metadata?

Abc <2134>xyz</2134>

No, the <2134> symbol should mean something.

The metadata should allow further interpretation.

What is metadata about?

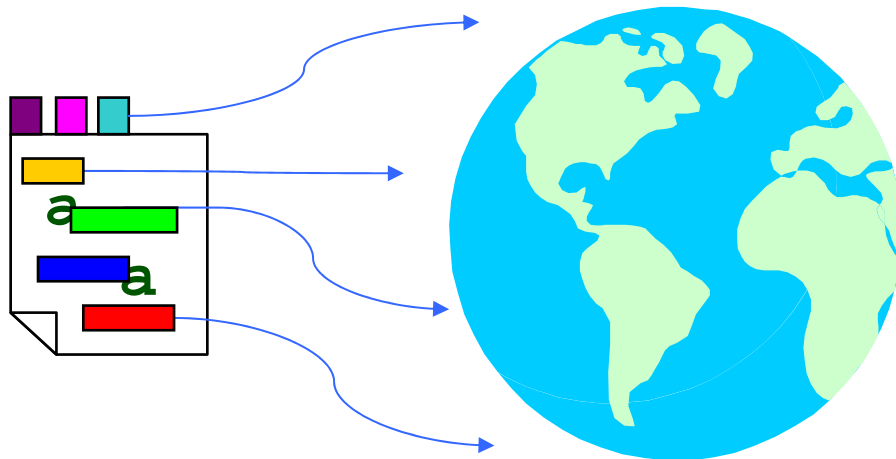


About RDF(S)?

About Knowledge
Engineering?

About research projects?

About socio-economic
impact?



The web is about the world.

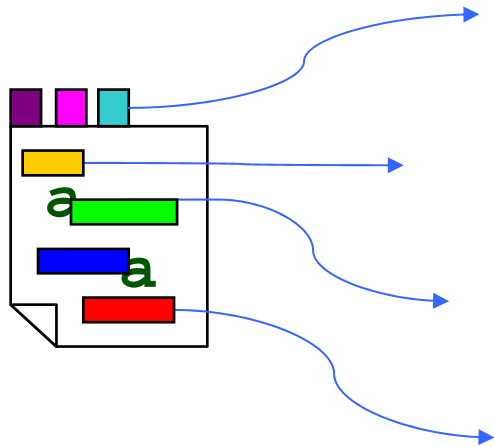
Directly: states, qualities,
intentions, descriptions

Or Indirectly: ideas, concepts,
documents.

Then the metadata should also
be about the World.

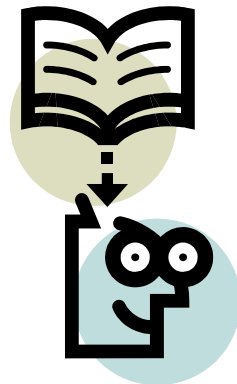
Not just symbols!

**Symbols referring
something in the world.**



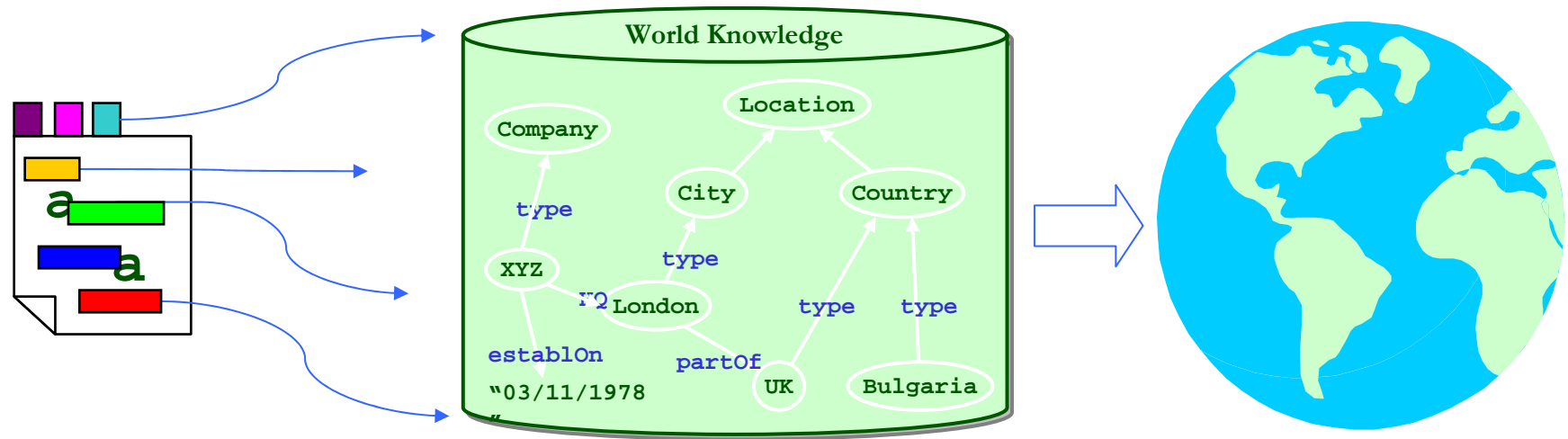
About specific things:

- Entities, Events
- Numbers, dates, currencies, documents



About general things:

- Concepts
- Relations
- Words



To enable metadata generation an usage there should be:

- **Model and methodology** for modelling world knowledge
- **World Knowledge** which allows further interpretation

- Introduction
 - Information access methods
 - Semantic Web and Metadata
 - **Our Approach**
 - Information Extraction
 - Semantic Annotation
 - Indexing & Retrieval
 - Model and Representation
 - KIM Platform: Implementing the Vision
-

- Provide **metadata schema**
 - Focusing on **entities (particulars)**
 - Define specific **tasks for generation and usage** of the metadata, which are
 - Well understood and measurable
 - Why not metadata about general things (universals)?
 - It is too complex
 - But we leave the door open
 - The particulars seem to provide a good **80/20 compromise**
-

Recent large scale human interaction study on a personal content IR system of Microsoft ([10]) demonstrates 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.

- The Semantic Annotation approach we offer is inspired by the Information Extraction
- We see **Information Extraction** (IE) as an enabling technology, because
- It offers a path for automatic metadata generation
- In robust, scalable, and predictable way
- And there is technology, which can **deliver this today**

- IE is a young discipline in NLP, which conducts partial analysis of text in order to extract specific information
- **Philosophy:** process bottom-up with a specific goals
- Concentrate on guaranteeing some level of performance
 - make simple things, but
 - In predictable and reliable fashion
 - With controlled accuracy (precision and recall)



There is a name for this specific things, used in NLP:

Named entities are considered: *people, organizations, locations*, and others referred by name. Can include also scalar values: *numbers, addresses, amounts of money*, etc. (NUMEX, TIMEX)

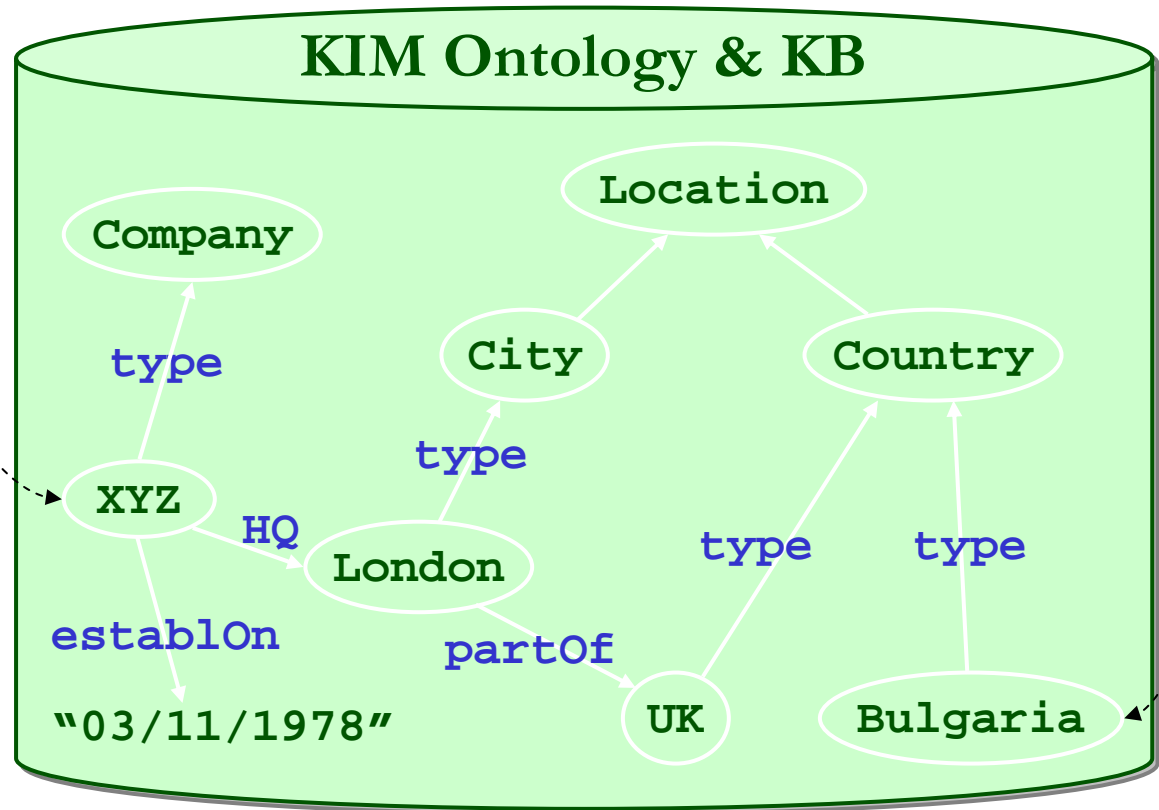
- The named entities and the words have different semantics
 - the first denote particulars (individuals),
 - the second – universals (concepts, classes, relations, attributes).
- So, while the words require handling of lexical semantics and common sense, the understanding and management of named entities, requires more specific world knowledge.

Typical tasks (see MUC and ACE):

- Named Entity extraction (including NUMEX, TIMEX)
 - Recognize People, Organizations, Dates, etc. in text
 - Attribute Extraction
 - Finding: *the colour of the car; the gender of a person*
 - Relation Extraction
 - extract that: *Mary is mother of Ivan; Ontotext is located in Sofia*
 - Scenario/Event Analysis
 - Understand the *time, place, and participants of a meeting*
-

- Introduction
 - Information access methods
 - Semantic Web and Metadata
 - Our Approach
 - Information Extraction
 - **Semantic Annotation**
 - Indexing & Retrieval
 - Model and Representation
 - KIM Platform: Implementing the Vision
-

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



- **Nature:** A specific metadata generation and usage schema
- **Motivating Hypothesis:** the named entities referred in the documents constitute important part of their semantics
- Two basic operations:
 1. **Annotate and hyperlink named entities** in documents;
 2. **Index and retrieve documents** wrt the referred entities.
- “Side effects”
 - **Ontology population** (Knowledge Acquisition)
 - When there are unknown entities in the text
 - Or unknown/new/changed attributes and relations

- This task can be seen as a combination of:
 - basic **press-clipping** exercise,
 - typical IE task, and
 - **automatic hyper-linking**.
- The resulting annotations represent:
 - a **document enrichment and presentation method**
 - Which can further be used to **enable other access methods**.

<i>John</i>	<i>loves</i>	<i>Mary</i>	
N	V	N	Basic NLP
Person		Person	IE system off the self
	love-v1		Semantic Tagging (WSD)
<u>http://..kim/Person457</u>	<u>http://..kim/Situat931</u>	<u>http://..kim/Person931</u>	Semantic Annotation (KIM)

- Semantic Annotation is about **assigning to the entities in the text links** allowing further formal interpretation.
- Provides **class and instance references** about the entities.
- It is a matter of terminology whether these annotations should be called “semantic”, “entity” or some other way
- It enables:
 - Highlighting, indexing and retrieval, categorization,
 - Generation of more advanced metadata
 - Smooth traversal between text and relevant knowledge

- Just a modification of the classical IR task:
 - documents to be **indexed according to NEs**;
 - retrieved based on **relevance to NEs**.
 - Basic assumption: the documents are characterized by the **bag of tokens** of their content, disregarding its structure.
 - In **IR typically**: tokens = word stems
 - Recently development towards using **word-senses** or lexical concepts for indexing and retrieval.
 - The named entities can be seen as special sort of tokens.
 - We present one more (pretty much independent) development direction instead of alternative of the contemporary IR trends.
-

- **Alias independent indexing:** index both “NY” and “N.Y.” as occurrence of the specific entity “New York”.
- **Advanced semantic querying** becomes feasible.
 - In a query, it is possible to specify entity type restrictions, name, and other attribute restrictions, as well as relations between the entities of interest.
 - For instance, query that targets all documents that refer to **Persons that hold some Positions within an Organization**
- **Range searches:**
 - Simplistic: give me documents referring number in the **range 13-14 mill**, or a date with the **1990s**

Information Need Definition:

1. Entity lookup, “give me a City with alias NY”
2. Entity pattern search, “give me CEOs of telecoms in Europe”
3. Combination, may include also keywords

Information need satisfaction:

1. Entities, or entity tuples (like <Person, Company, Location>)
2. Documents referring to entities or tuples

Can be seen as **formal factoid Question Answering**.

- Introduction
 - Information access methods
 - Semantic Web and Metadata
 - Our Approach
 - Information Extraction
 - Semantic Annotation
 - Indexing & Retrieval
 - **Model and Representation**
 - KIM Platform: Implementing the Vision
-

There are number of basic prerequisite for representation of semantic annotations:

- **Ontology** (or at least taxonomy) bearing the classes of entities.
- **Unique entity identifiers** which allow, those to be identified and linked to their semantic descriptions;
- **Knowledge base/repository** with entity descriptions.

- Two approaches for managing metadata and particularly annotations:
 - Embedded within the text (as in SGML, HTML, etc.)
 - Kept separate from the text (TIPSTER, GATE, Open Hypermedia Systems)
 - Number of pros and cons in general.
 - For the Semantic web, **non-embedded semantic annotations** seem better, because of complexity
 - Non-embedded semantic annotations allow **contextualized/customized/personalized metadata**
 - Which just means **Dynamic Semantic Web!**
-

Keep 'em separated



Keep separate:

- the content
- the metadata (annotations)
- the world knowledge

Three good reasons:

1. Could be created and owned separately
2. Could be modified separately
3. Could be too complex and inefficient to manage them together

- Introduction
 - Information access methods
 - Semantic Web and Metadata
 - Our Approach
 - Information Extraction
 - Semantic Annotation
 - Indexing & Retrieval
 - Model and Representation
 - **KIM Platform: Implementing the Vision**
-

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

A lot of the world knowledge is shared it is basic and shared or **common knowledge**, based on social, cultural, historical, and education context. Two types:

- **Common sense**, knowing that:
 - apple is tangible, but can't act as a dog, person and company can;
 - its more probable for a fog to spill rather than a skyscraper;
 - the number 3 can not be located somewhere, but car and meeting could;
 - **Common culture**, such as:
 - Events, people, numbers (Pi), dates
 - Films, actors, artists, sculptures, songs
 - Political leaders and parties
 - Sport clubs, famous players
 - Companies, famous people,
-

Number of “minor” obstacles:

- Modelling this sort of knowledge is a well-known *tough* philosophical, cognitive, and AI problem:
 - Cyc demonstrates how complex it is to model even a small portion of it;
 - The common part of it is quite fuzzy
 - One can expect that people from one district, school and social status, share quite a lot, but ... still not everything.
 - How to determine what’s common between CS student from Seattle and ballerina from Leningrad?
 - It is dynamic
 - Our grand children will probably not know who Mike Tison is.
 - It is not 100% language independent
 - How you think, depends on how you verbalize; no clear border between “world” and “lexical” concepts
-

Rationale:

- provide common knowledge about **entities**
- **KIM bets on scale** and avoids heavy semantics
- minimum common-sense, almost no axioms
- RDF(S) is being used in a way allowing migration to OWL (no multiple levels of meta-classes, etc.)

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

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

KIM “knows”:

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

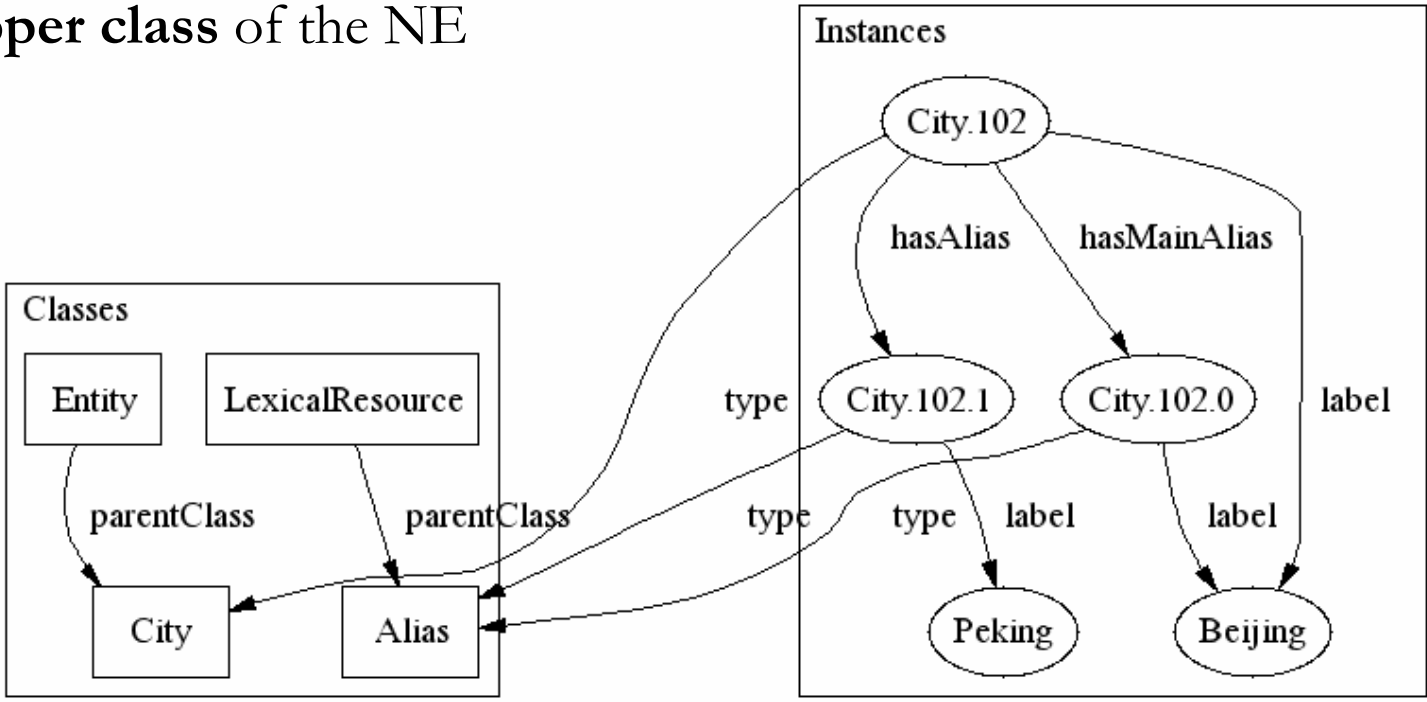
At present: about **200 000 pre-populated entities**:
50 000 locations, 130 000 organizations, 6000 people, etc.

Collected from various sources like **geographical** and
business intelligence gazetteers.

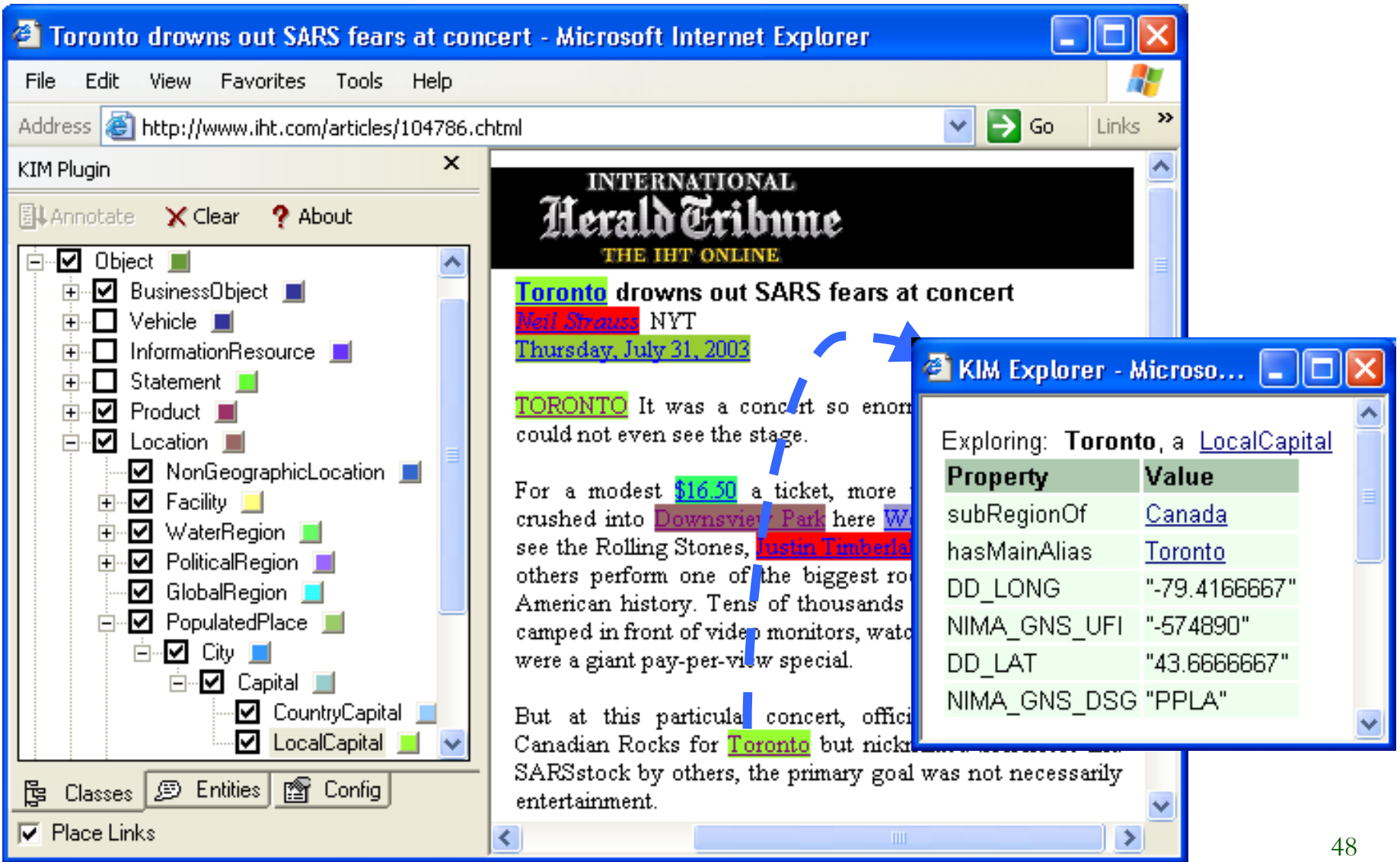
So, it is all predefined ... NO, KIM learns from the
texts it analyses

The NEs are described in terms of:

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



Usage? – Highlight and Explore

Toronto drowns out SARS fears at concert - Microsoft Internet Explorer

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

KIM Plugin

Annotations: Annotate, Clear, About

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

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.

KIM Explorer - Microsoft Internet Explorer

Exploring: Toronto, a LocalCapital

Property	Value
subRegionOf	Canada
hasMainAlias	Toronto
DD_LONG	"-79.41666667"
NIMA_GNS_UFI	"-574890"
DD_LAT	"43.66666667"
NIMA_GNS_DSG	"PPLA"

Entity Pattern Search

Search 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:

<input type="text" value="Z"/>	<input type="text" value="---"/>	<input type="text" value="is unknown"/>	<input type="text"/>
<input type="text" value="X"/>	<input type="text" value="---"/>	<input type="text" value="is unknown"/>	<input type="text"/>
<input type="text" value="Y"/>	<input type="text" value="---"/>	<input type="text" value="is unknown"/>	<input type="text"/>

Interested in



Entity Pattern Search



[Datstore](#) [Entity Pattern Search](#) [Predefined Patterns](#) [Entity Lookup](#)
[Keyword Search](#)

Document Query Result

Date	Title
15/10/2003 22:48	LG recruits U.S. fund in bid for South Korea's Hanaro

1-1 of 1 Documents per page: 15



Datastore	Entity Pattern Search	Predefined Patterns	Entity Lookup
Keyword Search			

Document Detail

Feature Name	Feature Value
TITLE	LG recruits U.S. fund in bid for South Korea's Hanaro
ORIGIN	SEOUL
SOURCE	IHT
UNIQUE_URL	http://www.iht.com/articles/113906.html

Document Content

SEOUL A battle between LG Group and a consortium led by American International Group for control of the No.2 South Korean high-speed Internet service provider intensified Wednesday after LG teamed up with a U.S.-\$ based investment fund to outbid its rival.

LG, a South Korean conglomerate, and Carlyle Group of Washington offered new shares valued at 736.2 billion won, or \$626 million, for a majority stake in Hanaro Telecom, in addition to arranging \$700 million through syndicated loans.

If their bid is successful, LG and the fund will manage Hanaro jointly, holding a 51 percent stake. LG already owns 18 percent of Hanaro.

AIG and Newbridge Capital, another global investment fund based in the United States, this year offered \$500 million for almost 40 percent of Hanaro. The company's shareholders

Give **KIM** a try:

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