Second European Business Intelligence Summer School (eBISS 2012)

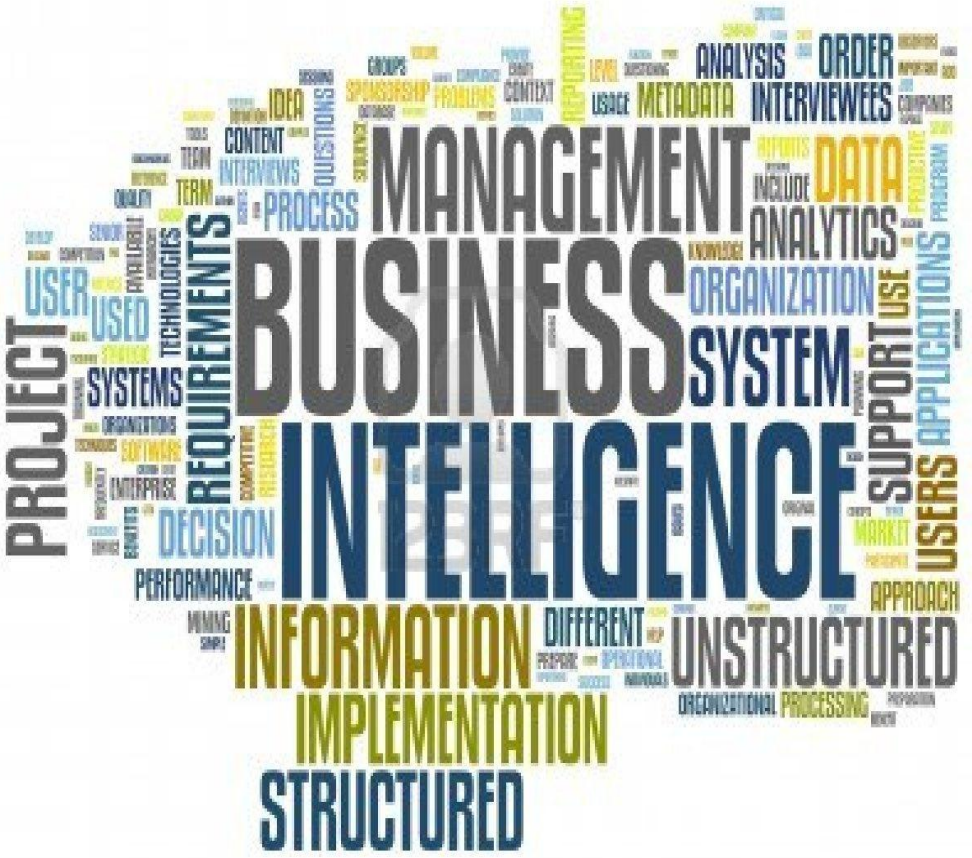
July 15 - 21, 2012 Brussels, Belgium

Knowledge Harvesting for Business Intelligence

Nesrine Ben Mustapha and Marie-Aude Aufaure

Nesrine.ben-mustapha@ecp.fr

CENTRALE PARIS


ANALYSIS
ORDER
METADATA
INTERVIEWEES
MANAGEMENT
DATA
ANALYTICS
BUSINESS
ORGANIZATION
SYSTEM
INTELLIGENCE
INFORMATION
UNSTRUCTURED
IMPLEMENTATION
STRUCTURED
PROJECT
REQUIREMENTS
USERS
SUPPORT
APPLICATIONS
DECISION
PERFORMANCE
DIFFERENT
APPROACH
PROCESS
USER
USED
SYSTEMS

Business

Context

Intelligence in Business Intelligence (BI)?

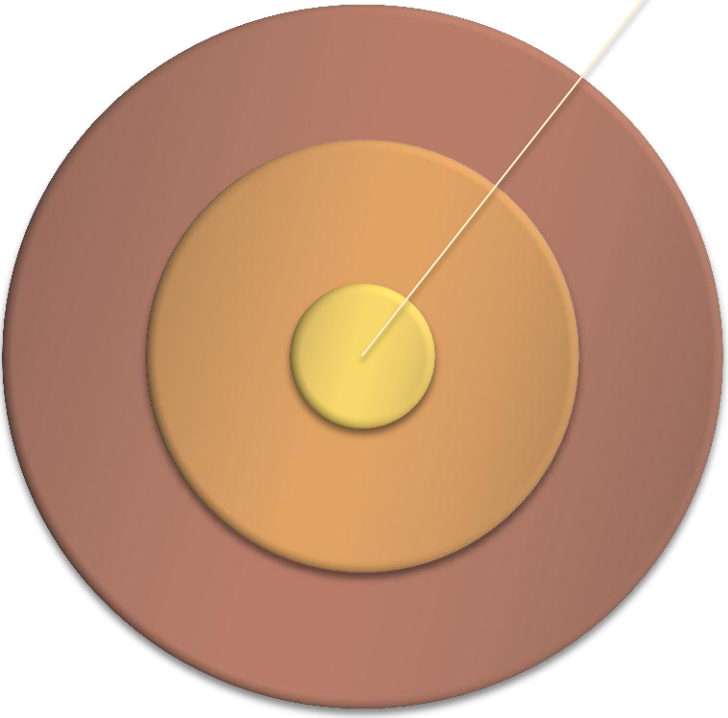
- Intelligence:
  - application of Information, skills, experiences and reasoning to solve business problems
  - Information acquisition from wide varity of sources
  - Harvesting knowledge for decision making

Talk Overview

- Context of BI: need of semantics
- Correlated dimensions related to semantic BI
  - Web evolution
  - Semantic evolution
  - New trend of search paradigm
  - Progress in ontology engineering
- Ontology-based knowledge harvesting for Business Intelligence

Context

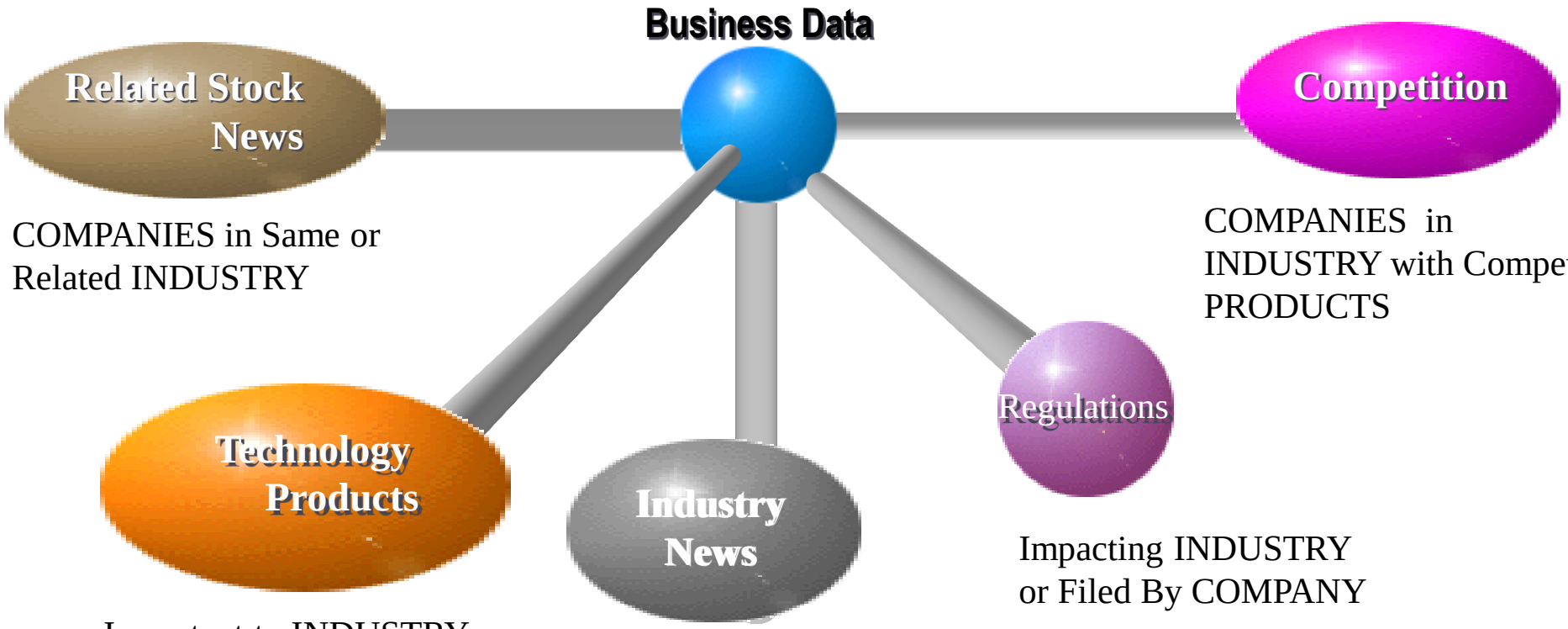
Data, Information and Knowledge in BI Process



Data

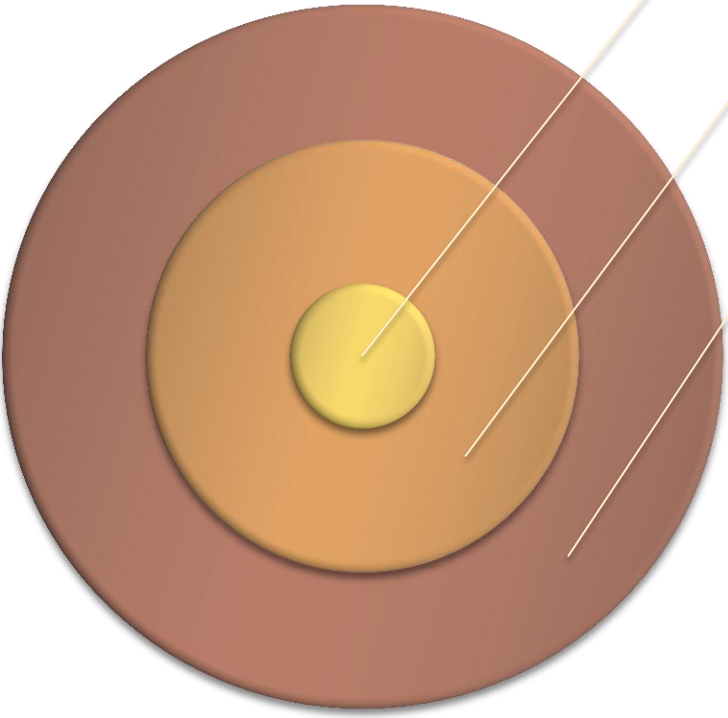
- What data is to be collected and managed and in what context ?

Context
Data of BI



Context

Data, Information and Knowledge in BI



Data

Information

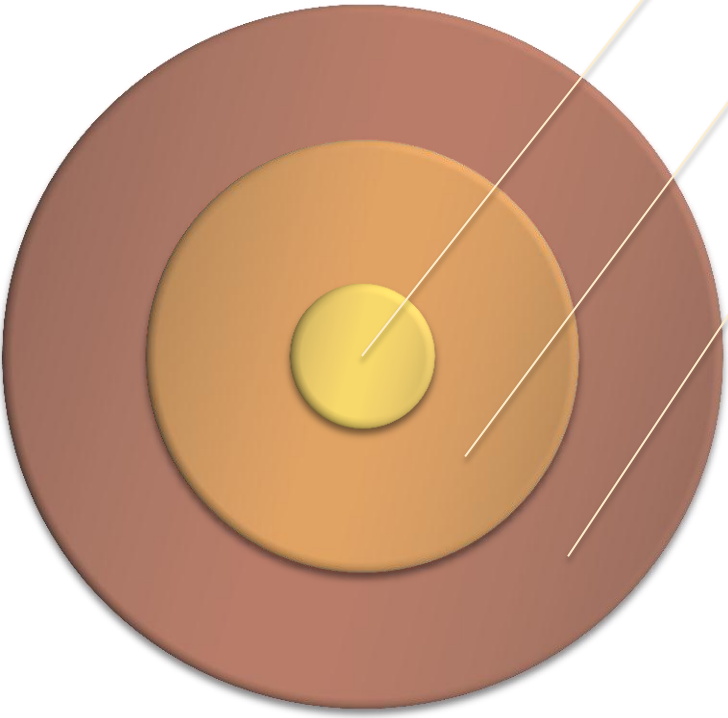
Knowledge

- What data is to be collected and managed and in what context ?

- Data warehousing, online analytical processing, data quality, data profiling, business rule analysis, and data mining

Context

Data, Information and Knowledge in BI



Data

- What data is to be collected and managed and in what context ?

Information

- Data warehousing, online analytical processing, data quality, data profiling, business rule analysis, and data mining

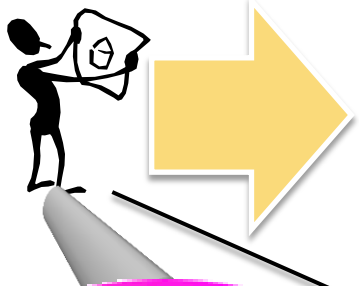
Knowledge

- Business plans
- Explicit, embedded or tacit Knowledge?

Business Intelligence: closed or open environment

BI Environment

Knowledge Exploiters



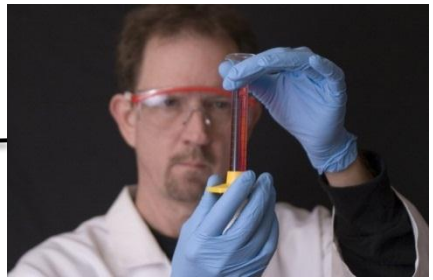
Competition

Knowledge learners

May be focused on certain areas and not able to integrate different streams of knowledge.

Knowledge Explorers

Maintain a good balance between internal and external learning.



Knowledge Innovators

Combine internal and external learning

Motivating Use cases
Business Case study in Knowledgeweb Network

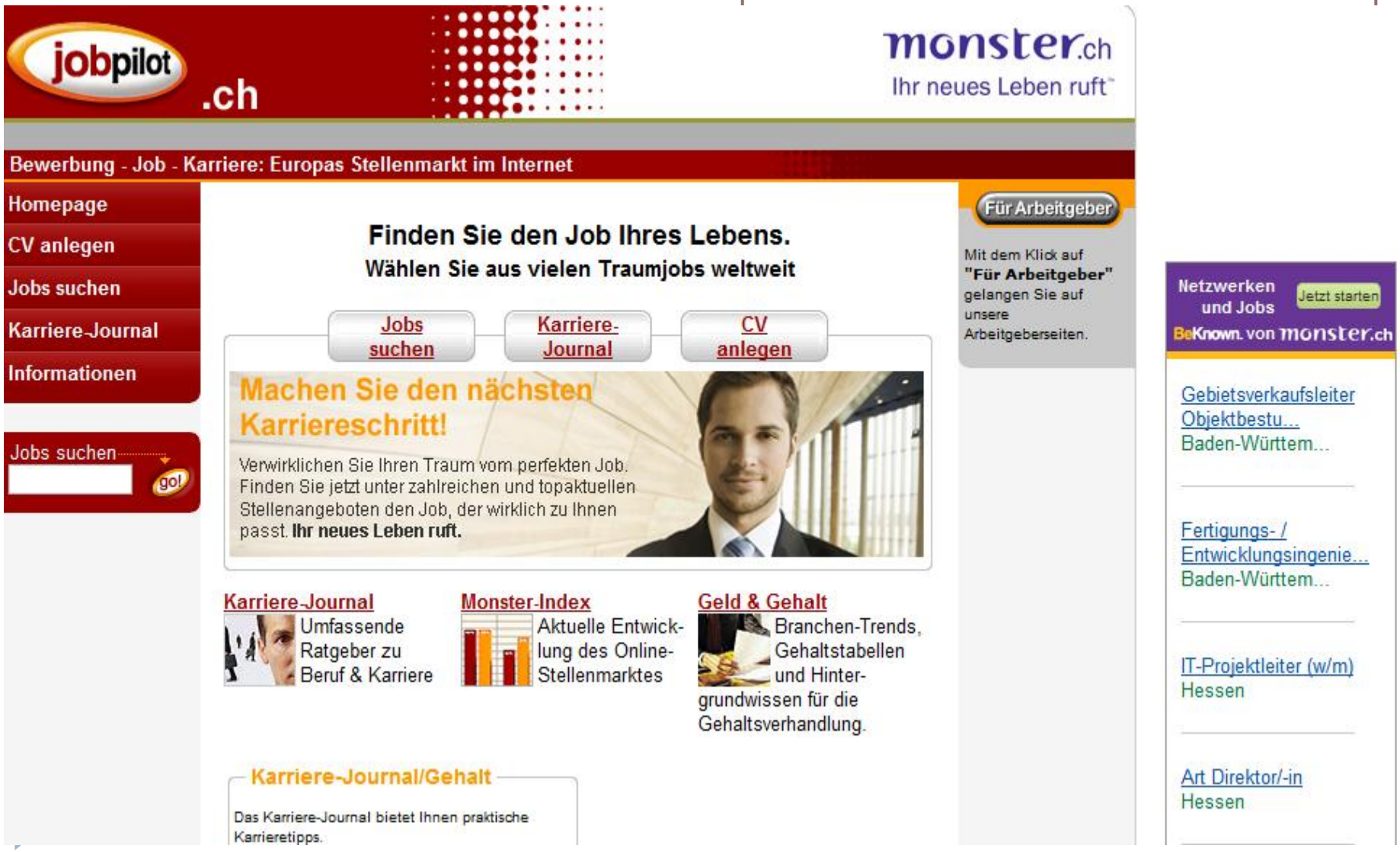
Main Goal: Transfer of ontology-based technologies from the field of academia to industry

- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media and comms**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

Motivating Use cases
Recruitment

State job Center

- German federal Employment office
- Monster



Motivating Use cases
Recruitment

- Challenges and new requirements
  - Facilitate efficiently open job vacancies with qualified suitable candidates
  - → Automatic matching between job offers and job seekers
- → Semantic Solution:
  - → Semantic support based on expressing relationships between job characteristics and candidate qualifications
    - → Representing
    - → Searching
    - → Sharing on the web
    - → Semantic matching

Motivating Use cases
Semantic solution and key business benefit

Job offers

Job seekers profiles

Solution

Semantic querying

Machmaker

Metadata crawler

Knowledge Base

Enriching

Semantics (ontologies)

Enriching

Motivating Use cases
Business Case study in Knowledgeweb Network

Main Goal: Transfer of ontology-based technologies from the field of academia to industry

- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media and comms**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

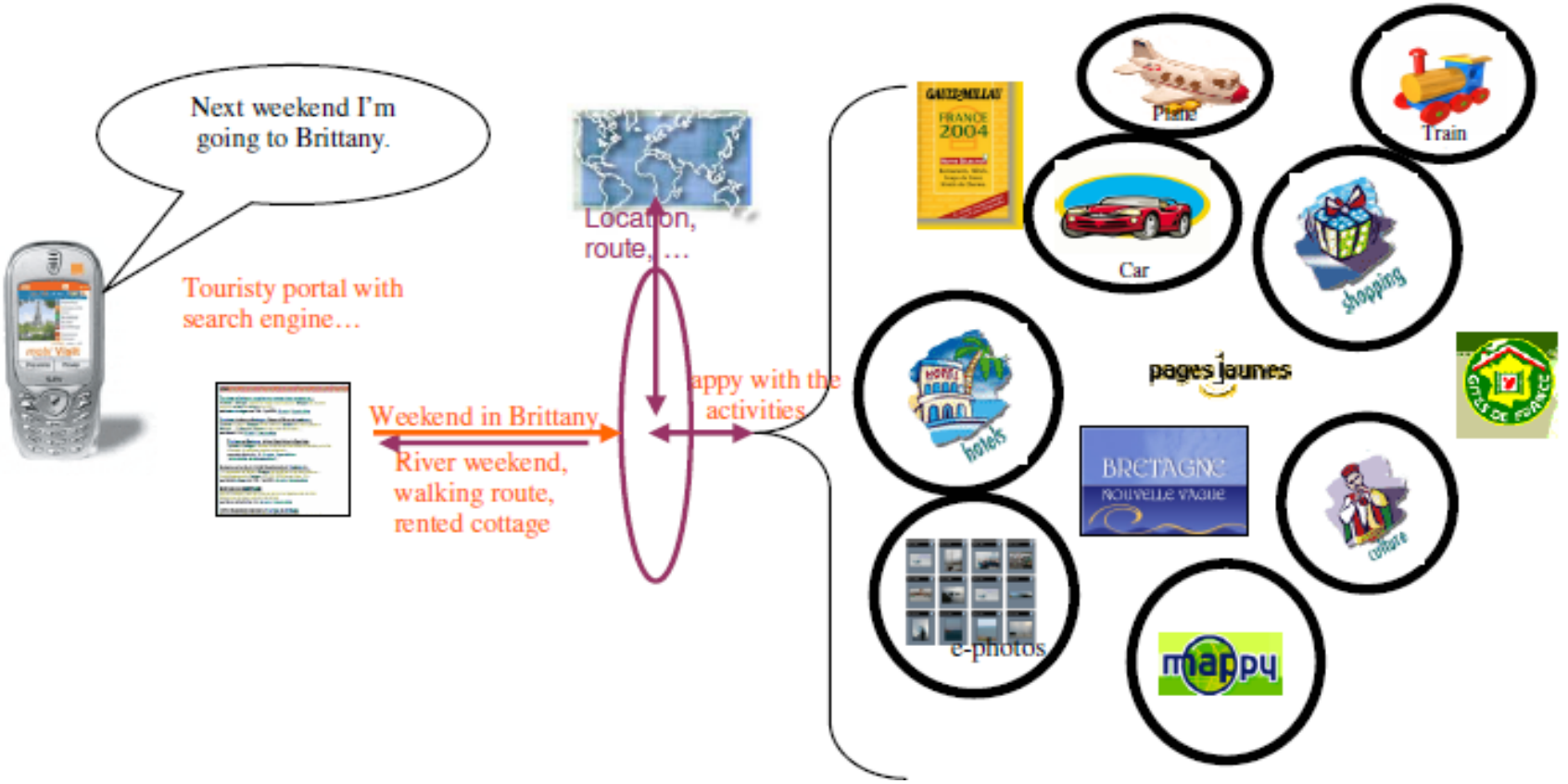
Motivating Use cases
B2C market place for tourism

- Current systems

Only information pages and no tourism offers
No personalized services

Motivating Use cases
B2C market place for tourism

- Challenges and new requirements
  - Offer on-line personalized tourism packages



Motivating Use cases
B2C market place for tourism

- Challenges and new requirements
  - Offer on-line personalized tourism packages
  - Geo-localization
  - Dynamic exploitation of content, service providers and personalized data
- Semantic Solution:
  - Semantic data integration
  - Natural language processing
  - Personal data representation and exploitation
  - Semantic web services

In France, Tourism market was evaluated 32 Billions euros which river tourism represents a turnover greater than 250 M euros.

Motivating Use cases
Business Case study in Knowledgeweb Network

Main Goal: Transfer of ontology-based technologies from the field of academia to industry

- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

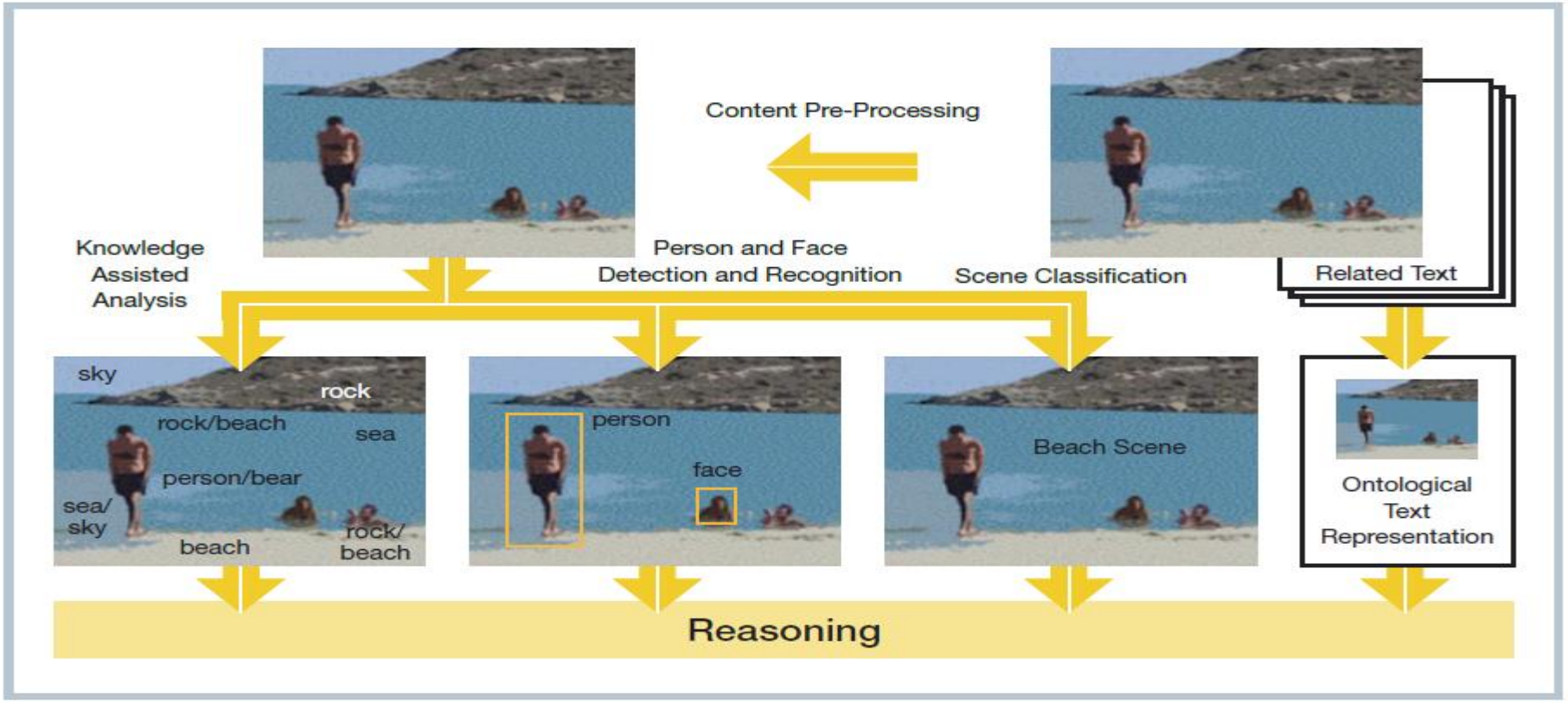
Motivating Use cases
Multimedia Content analysis

- Current systems
  - Difficult to develop and maintain large multimedia databases
  - Difficult to organize, find and distribute multimedia content
- Solutions
  - Multimedia data can be annotated in terms of knowledge extracted from it
  - Machine processable data models supporting
    - Semantic search
    - Navigation
    - Reasoning functionalities

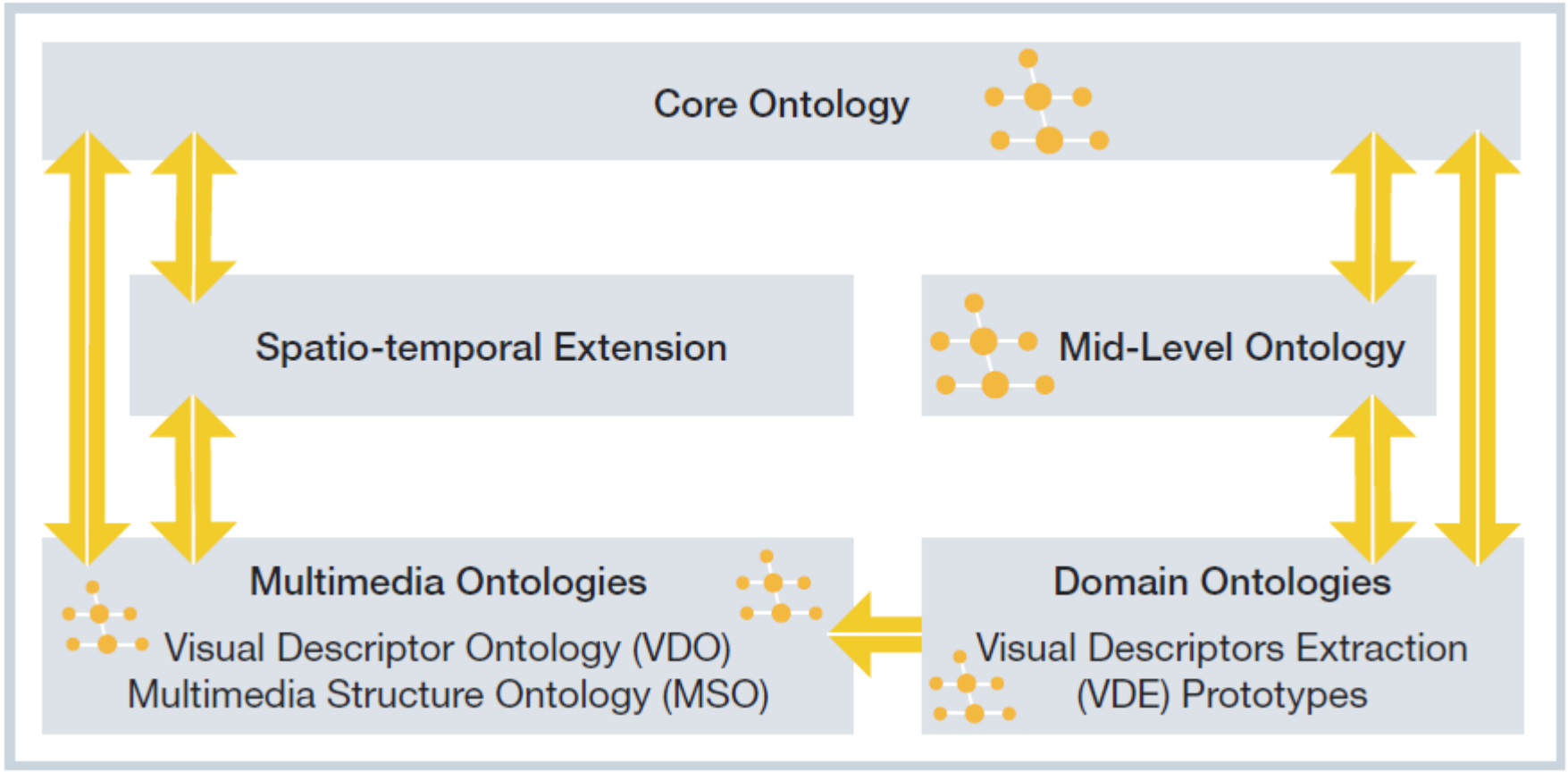
Motivating Use cases
Multimedia Content analysis

- AceMedia Project

Creating intelligent multimedia content



Motivating Use cases
Multimedia Content analysis



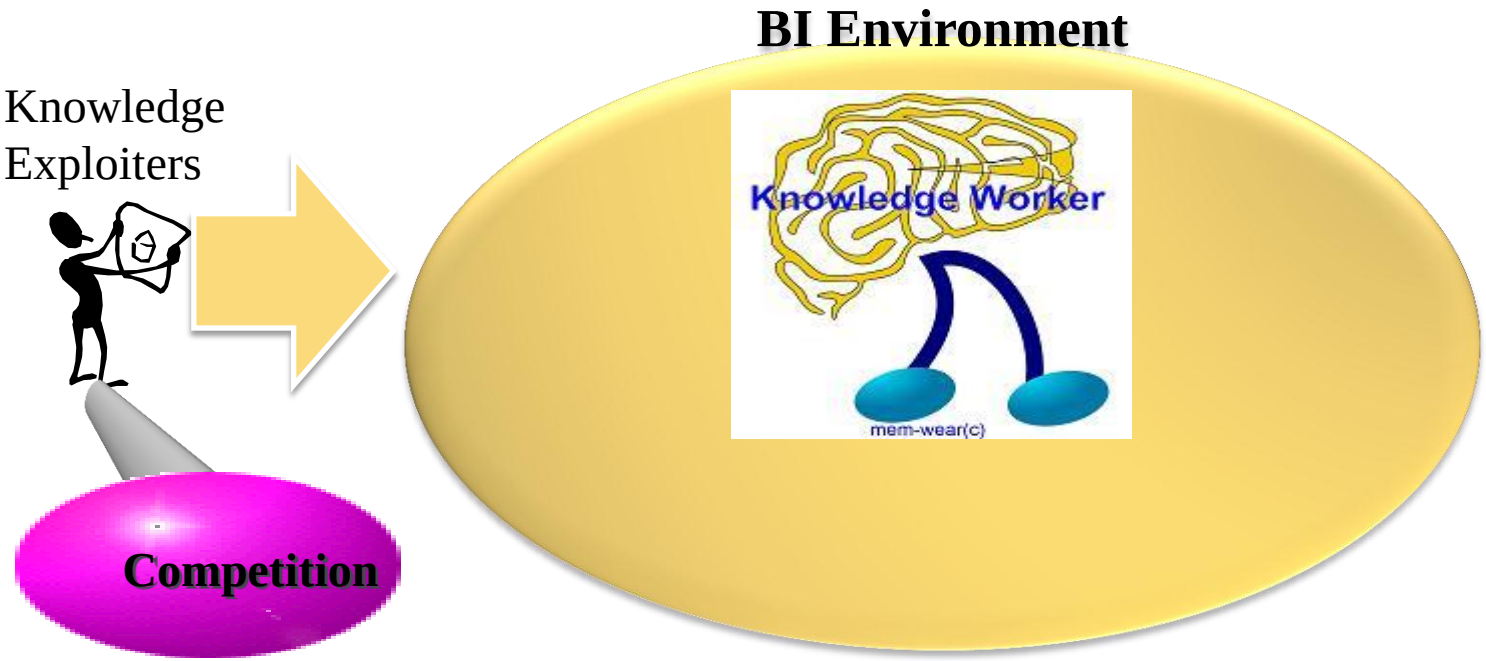
Motivating Use cases
Business Case study in Knowledgeweb Network

Main Goal: Transfer of ontology-based technologies from the field of academia to industry

- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

Motivating Use cases
News aggregation service

- Business interest in following news in specific categories including economics, sciences and IT, or on specific companies.
- Manual fastidious task !!

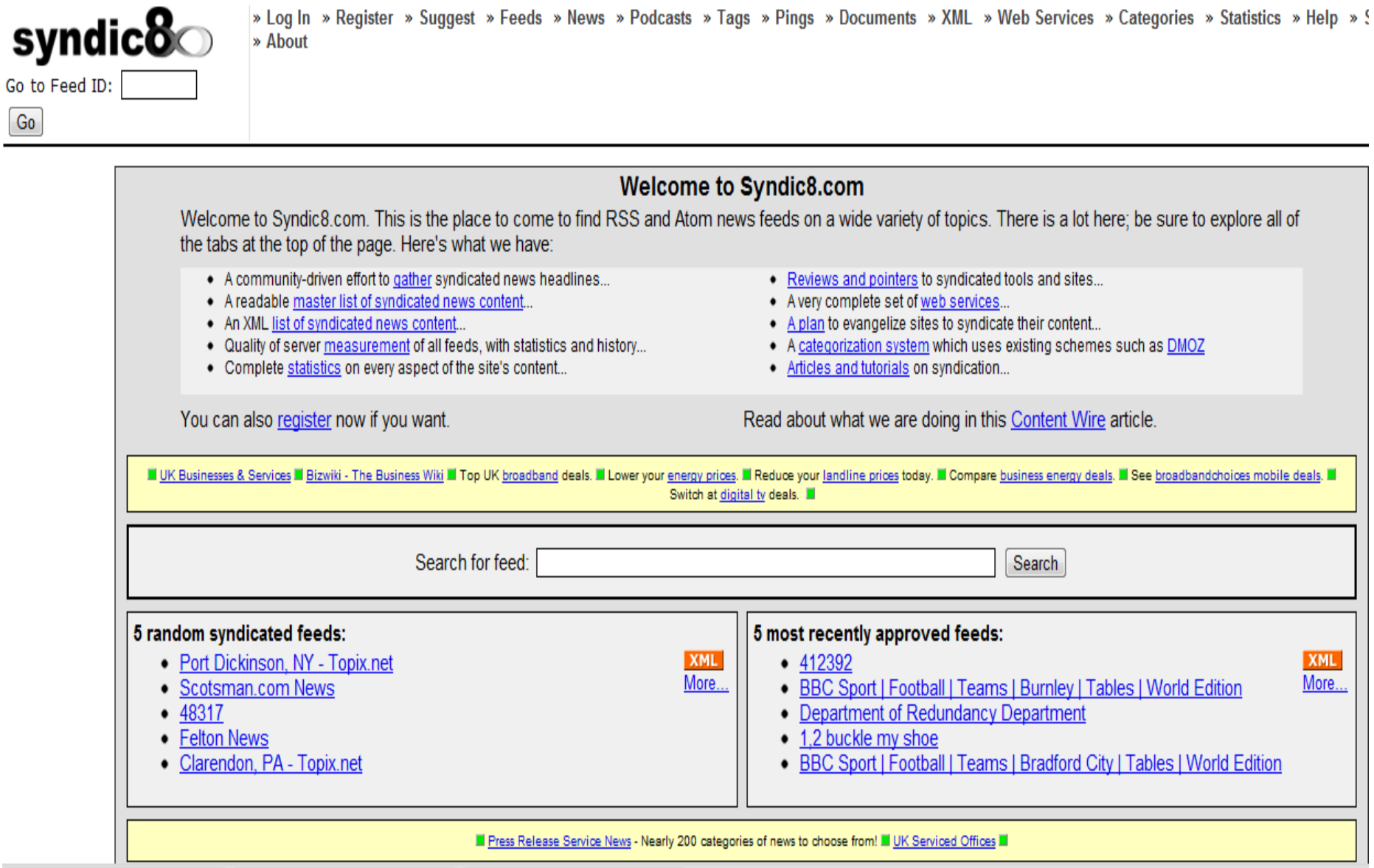


Motivating Use cases
News aggregation service

- Traditional systems:
  - Feed services based on RSS or Atom (www.Syndic8.com)

Motivating Use cases

News aggregation service



Motivating Use cases
News aggregation service

- Traditional systems:
  - Feed services based on RSS or Atom (www.Syndic8.com)
    - very basic Model (e.g. title, author, link to full story) and not suitable for any intelligent searching or organizing
  - Portals such as Google News
    - Large body of information that can be processed

Motivating Use cases
News aggregation service

- Current Semantic approaches
  - The news aggregation service http://www.newsexpress.de from neofonie GmbH

Motivating Use cases
News aggregation service

Motivating Use cases
News aggregation service

- Current Semantic approaches
  - The news aggregation service http://www.newsexpress.de from neofonie GmbH
  - Manual creation by a source expert of a XSLT template for each news source
  - Automatic processing of that news source through a thematic clustering algorithm (NLP) and classification with category mappings

Extraction of semantics from the source documents

Motivating Use cases
Data warehousing in healthcare

Main Goal: Transfer of ontology-based technologies from the field of academia to industry

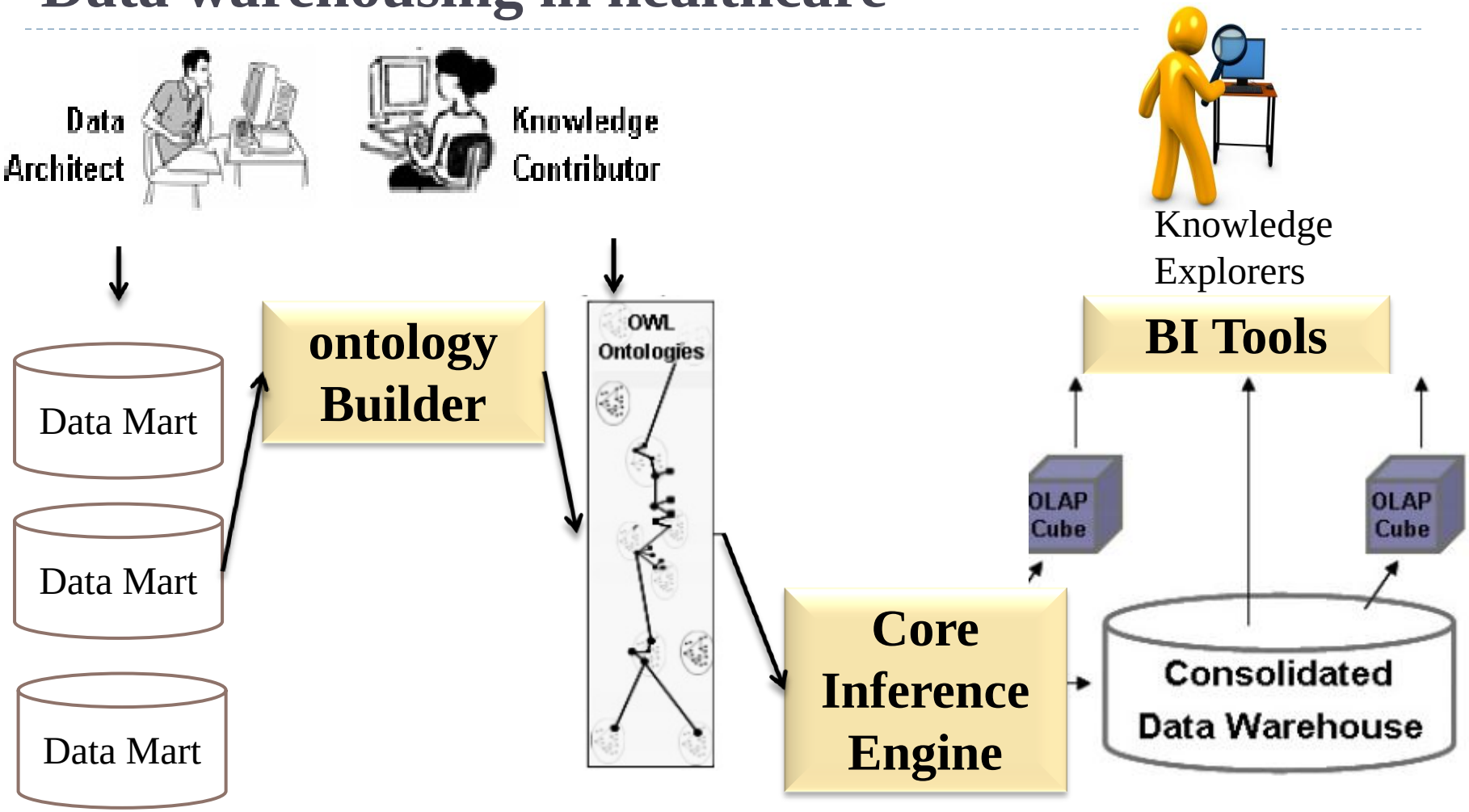
- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

Motivating Use cases
Data warehousing in healthcare

- Large health insurance company use a cognos data warehousing solution to administrate its data
- Business data are stored in various PC and don't share the same data formats
- Results:
  - Manual search over data sources
  - Need of introducing common terminology for healthcare data
  - Problem of updating data
- Semantic solution

Common terminology of healthcare domain for ontology-based integration

Motivating Use cases
Data warehousing in healthcare



Motivating Use cases
Data warehousing in healthcare

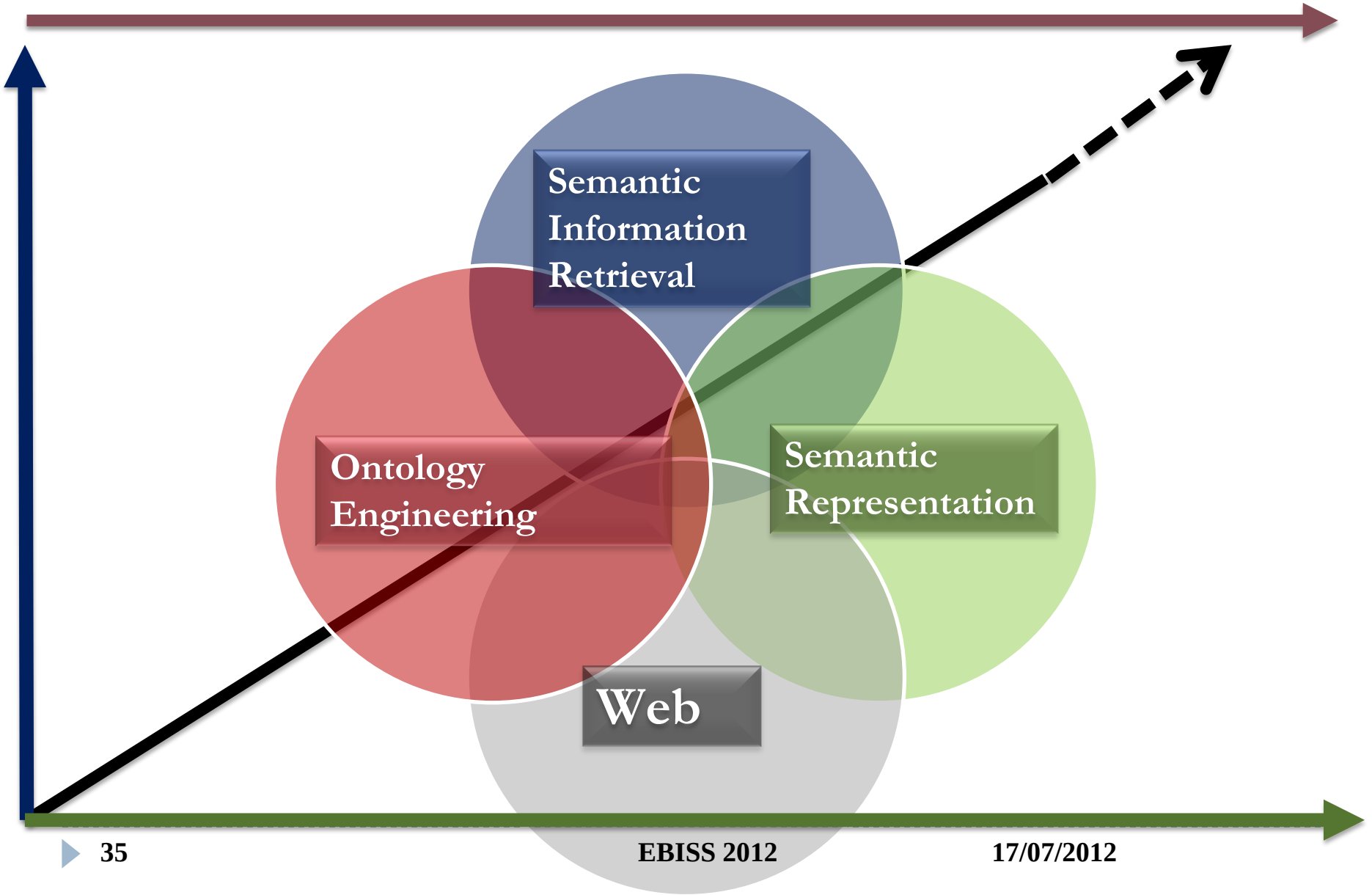
Main Goal: Transfer of ontology-based technologies from the field of academia to industry

- **Service Industry:**
  - **Recruitment**
  - **B2C market place for tourism**
- **Media**
  - **Multimedia Content analysis**
  - **News aggregation service**
- **Technology providers**
  - **Product lifecycle management**
- **Health:**
  - **Data warehousing in healthcare**
  - **Hospital information systems**

Motivating Use cases
Hospital information systems

- Hospitals have dispersed data sources: administrative information about patients, diagnoses and treatment history for each division
- Different type formats of stored data (databases, texts)
- Need of efficient access
- Semantic solution:
  - Ontology engineering from unstructured data
  - Data wrapper (from database to Ontology)
  - Query mediation and semantic matching solution
  - Middleware for database integration

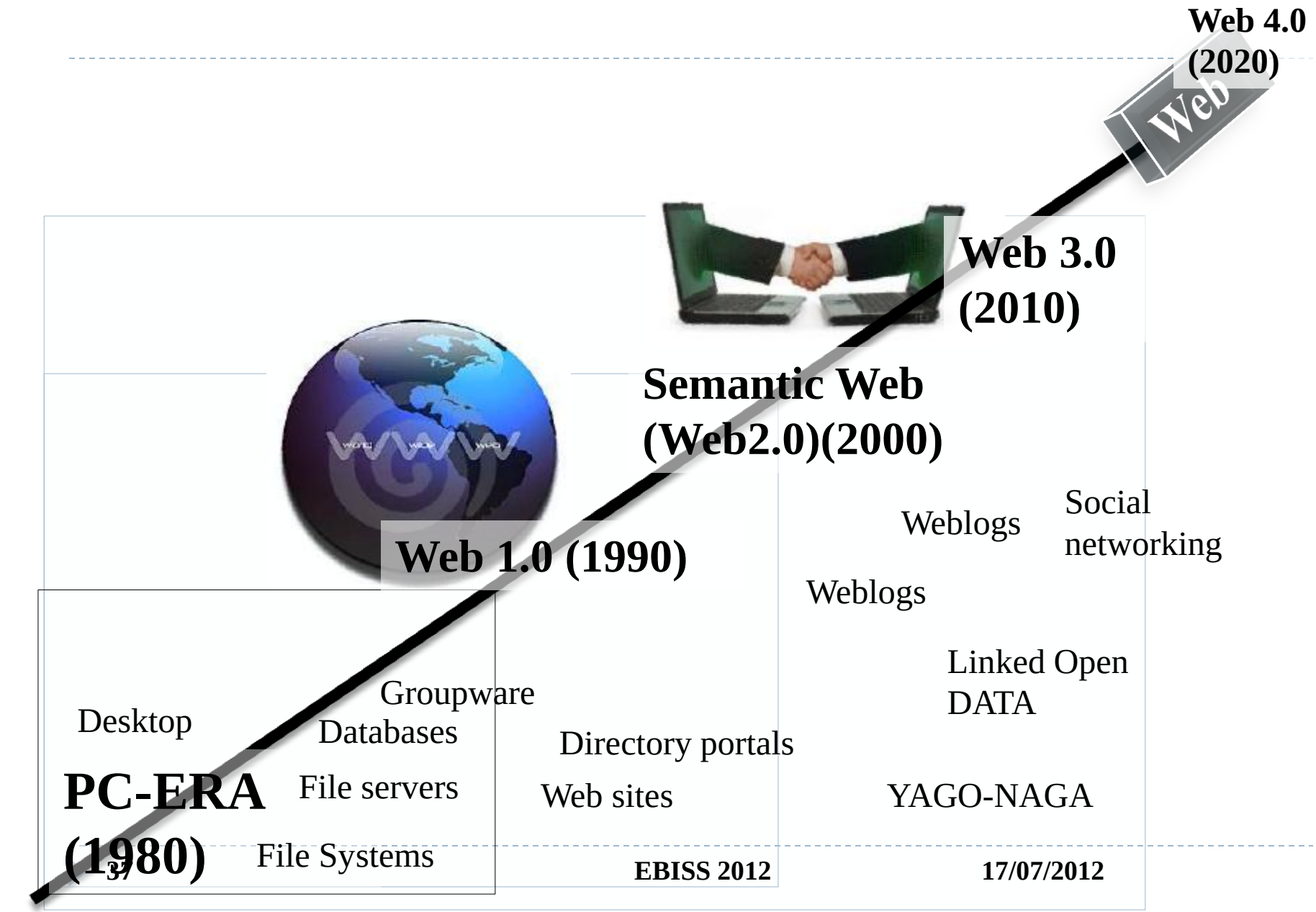
Correlated dimensions related to semantic BI



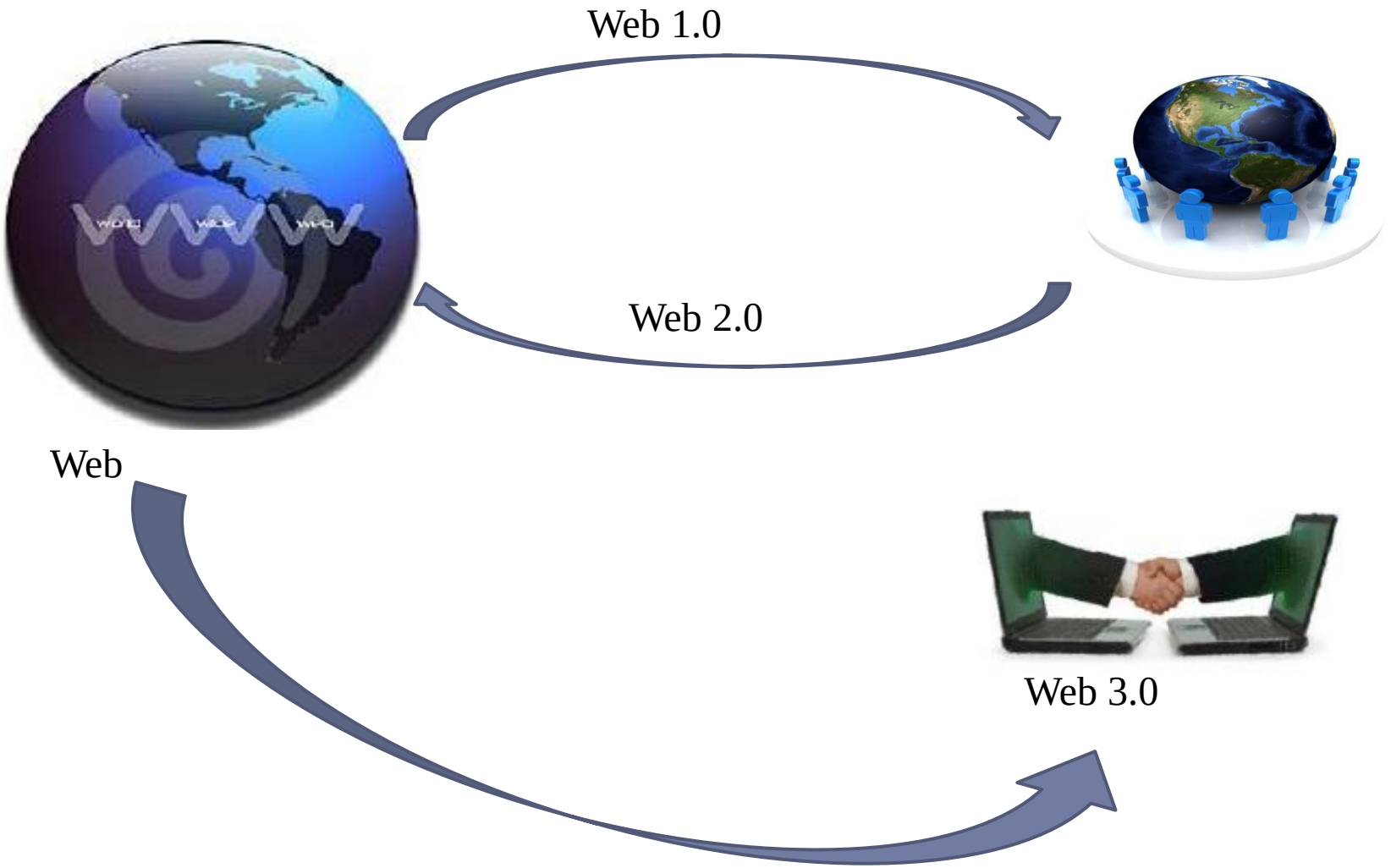
Evolution of Web: Semantic Web

Towards Open Linked Data

Evolution of the web



Evolution of the web
Semantic web



Evolution of the web

From Wikipedia to YAGO, DBPEDIA,…OLD

Corporation

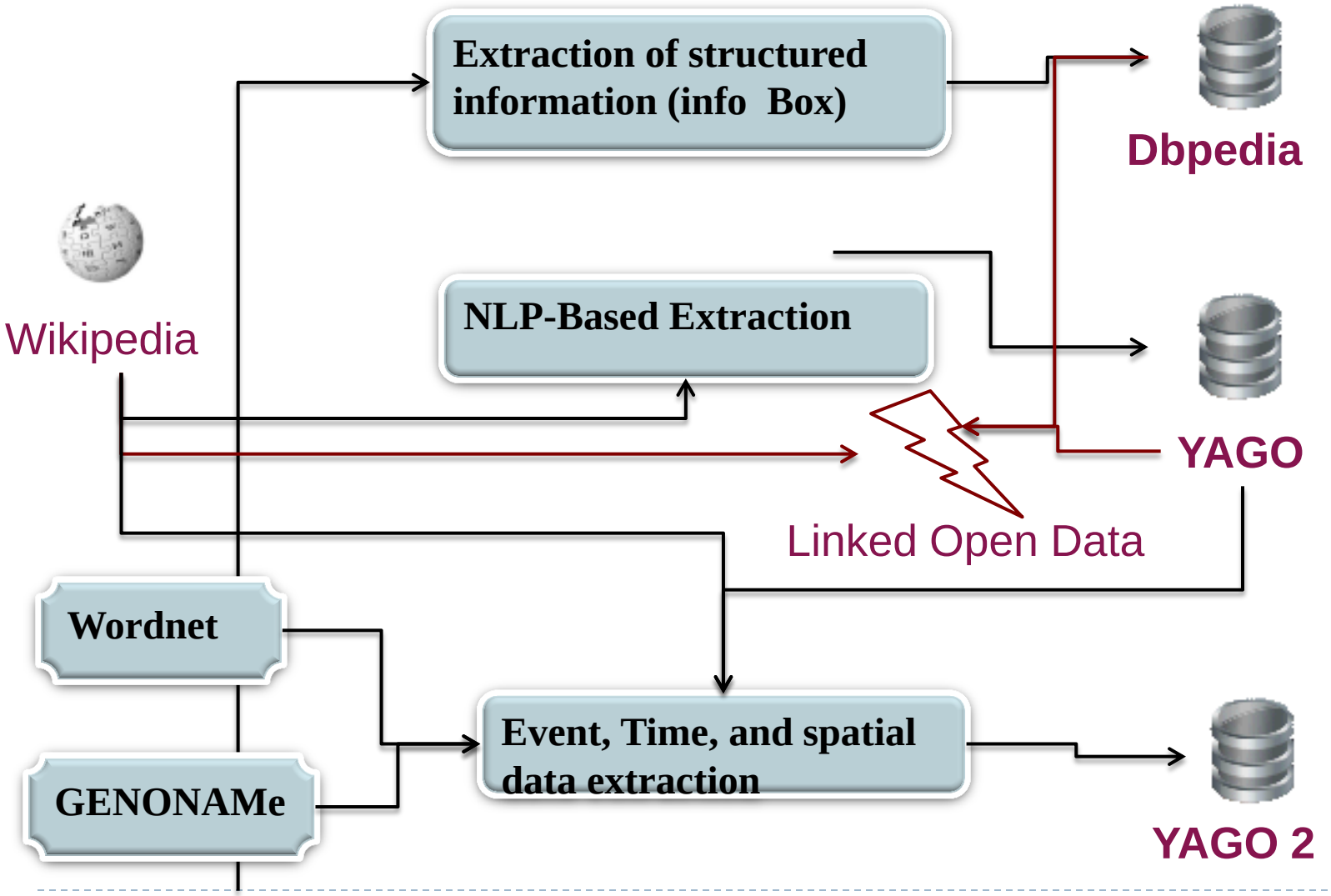
IBM

Logo designed by Paul Rand in 1972.

Type	Public
Traded as	NYSE: IBM Dow Jones Component S&P 500 Component
Industry	Computer hardware, Computer software, IT services, IT consulting
Founded	Endicott, New York, U.S. (June 16, 1911)
Founder(s)	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	Worldwide
Key people	Samuel Palmisano (Chairman) Ginni Rometty (President and CEO)
Products	See IBM products
Revenue	▲ US$ 106.91 billion (2011)[1]
Operating income	▲ US$ 20.28 billion (2011)[1]
Net income	▲ US$ 15.85 billion (2011)[1]

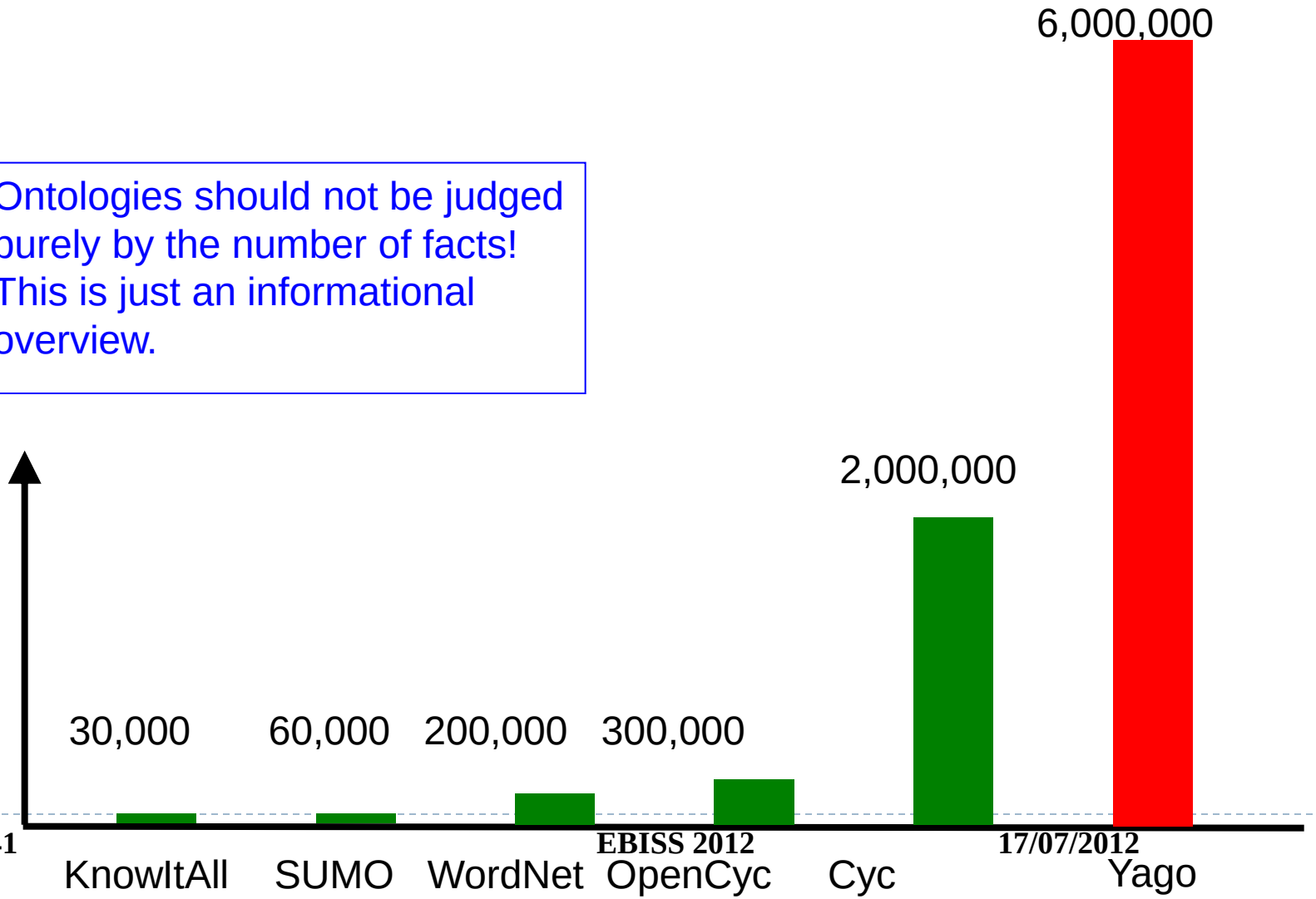
dbpedia-owl:formationDate	1911-06-16 (xsd:date)
dbpedia-owl:formationYear	1911-01-01 00:00:00 (xsd:date)
dbpedia-owl:foundationPlace	dbpedia:Endicott,_New_York
dbpedia-owl:foundedBy	dbpedia:Thomas_J._Watson
dbpedia-owl:industry	dbpedia:Information_technology
dbpedia-owl:keyPerson	dbpedia:Samuel_J._Palmisano dbpedia:President dbpedia:Chief_executive_officer dbpedia:Chairman
dbpedia-owl:locationCity	dbpedia:United_States dbpedia:Armonk,_New_York
dbpedia-owl:numberOfEmployees	426751 (xsd:integer)
dbpedia-owl:product	dbpedia:List_of_IBM_products
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/comm
dbpedia-owl:type	dbpedia:Public_company
dbpedia-owl:wikiPageExternalLink	http://source.icu-project.org/repos/icu/icu/tr http://www.ibm.com/ http://www-03.ibm.com/linux/project_big_gr http://asmarterplanet.com/ http://www.ibm.com/ibm/ideasfromibm/us/s http://query.nytimes.com/gst/fullpage.html?
dbpprop:areaServed	Worldwide
dbpprop:companyLogo	200 (xsd:integer)
dbpprop:companyName	International Business Machines
dbpprop:companyType	dbpedia:Public_company
dbpprop:description	Thomas J. Watson, who led IBM from 1914-
dbpprop:divisions	Software Services Hardware Financing
dbpprop:filename	Think Thomas J Watson Sr.ogg

Evolution of the web
Existing large knowledge bases

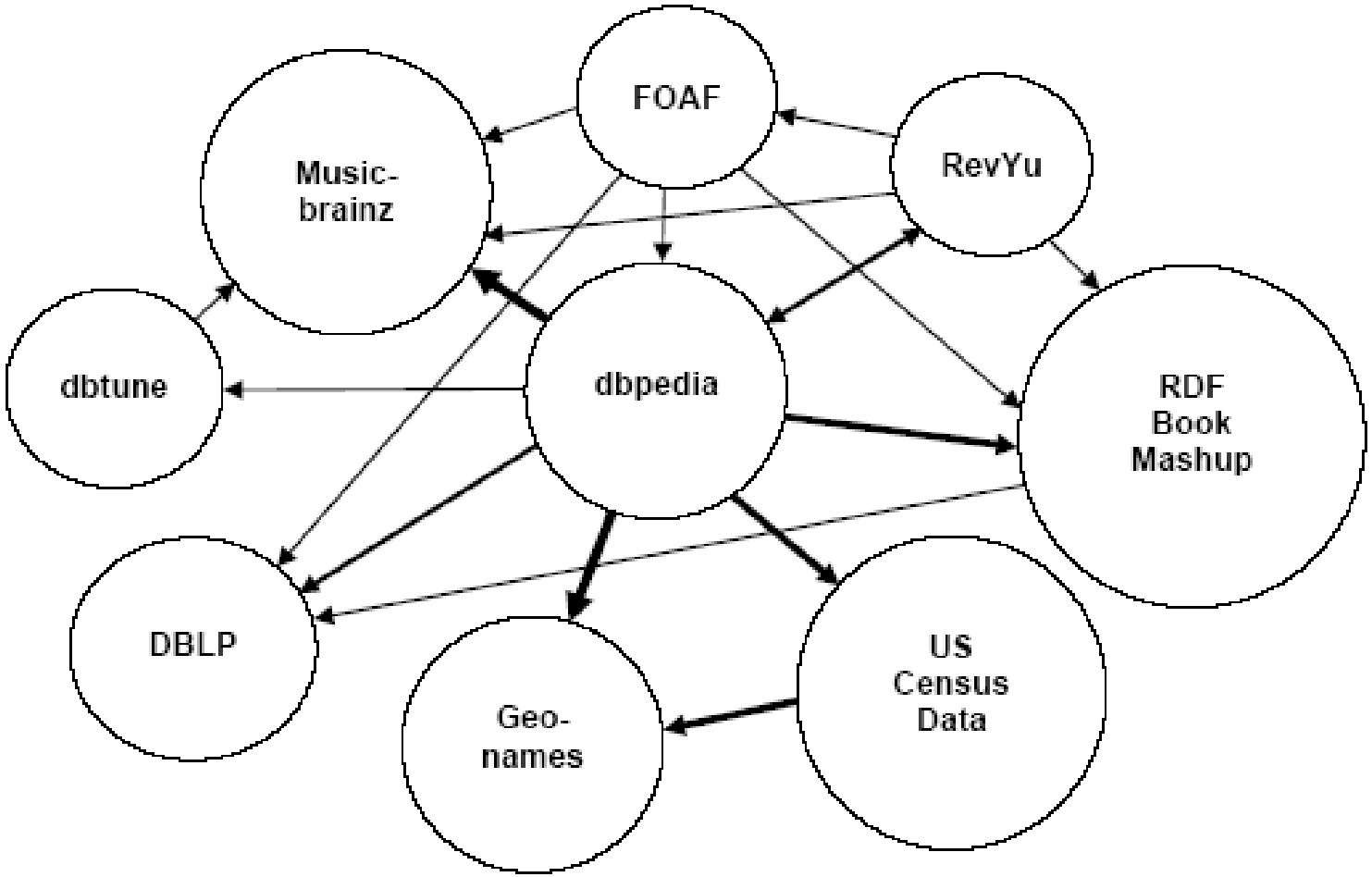


Evolution of the web
Existing large knowledge bases

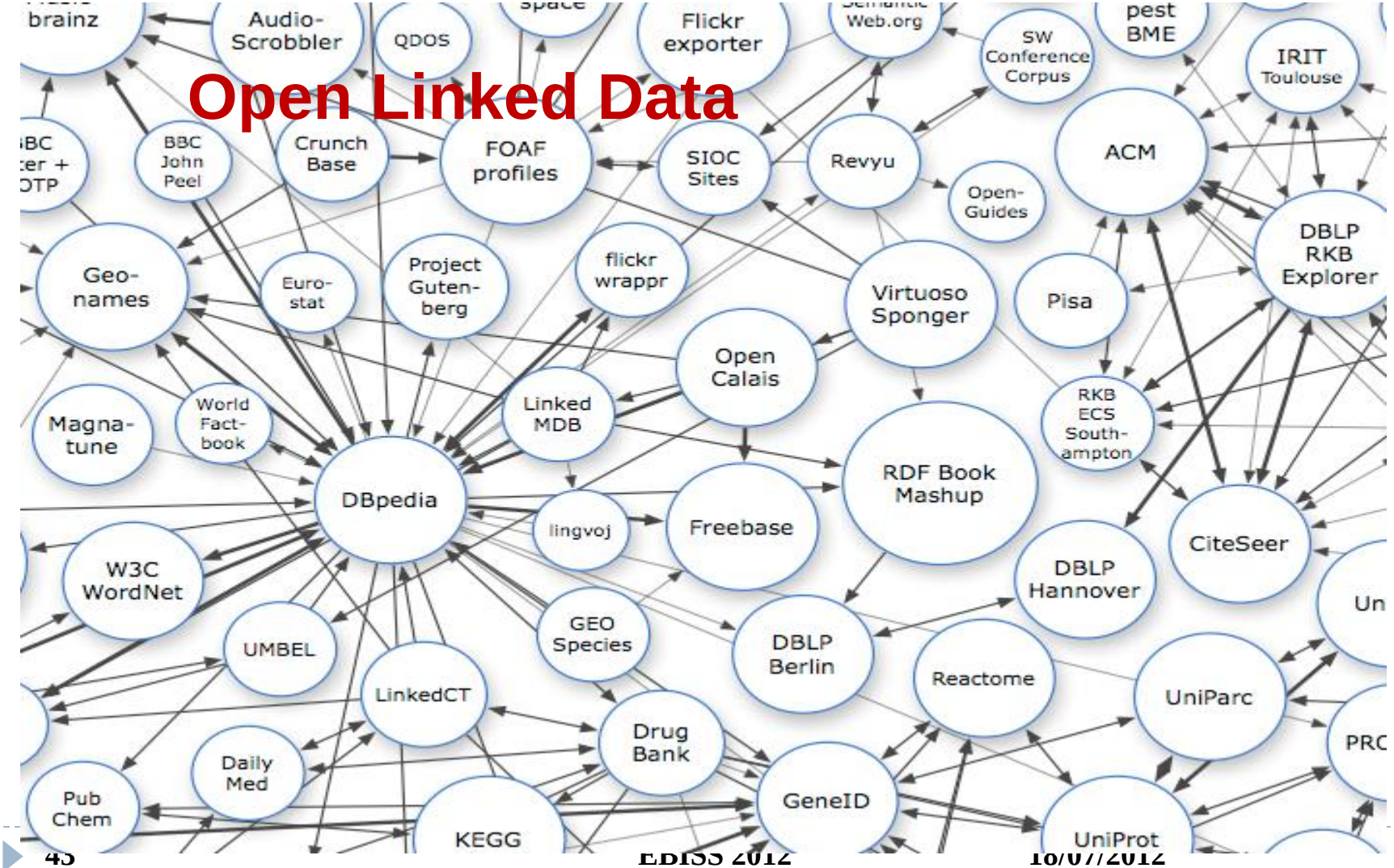
Ontologies should not be judged purely by the number of facts! This is just an informational overview.



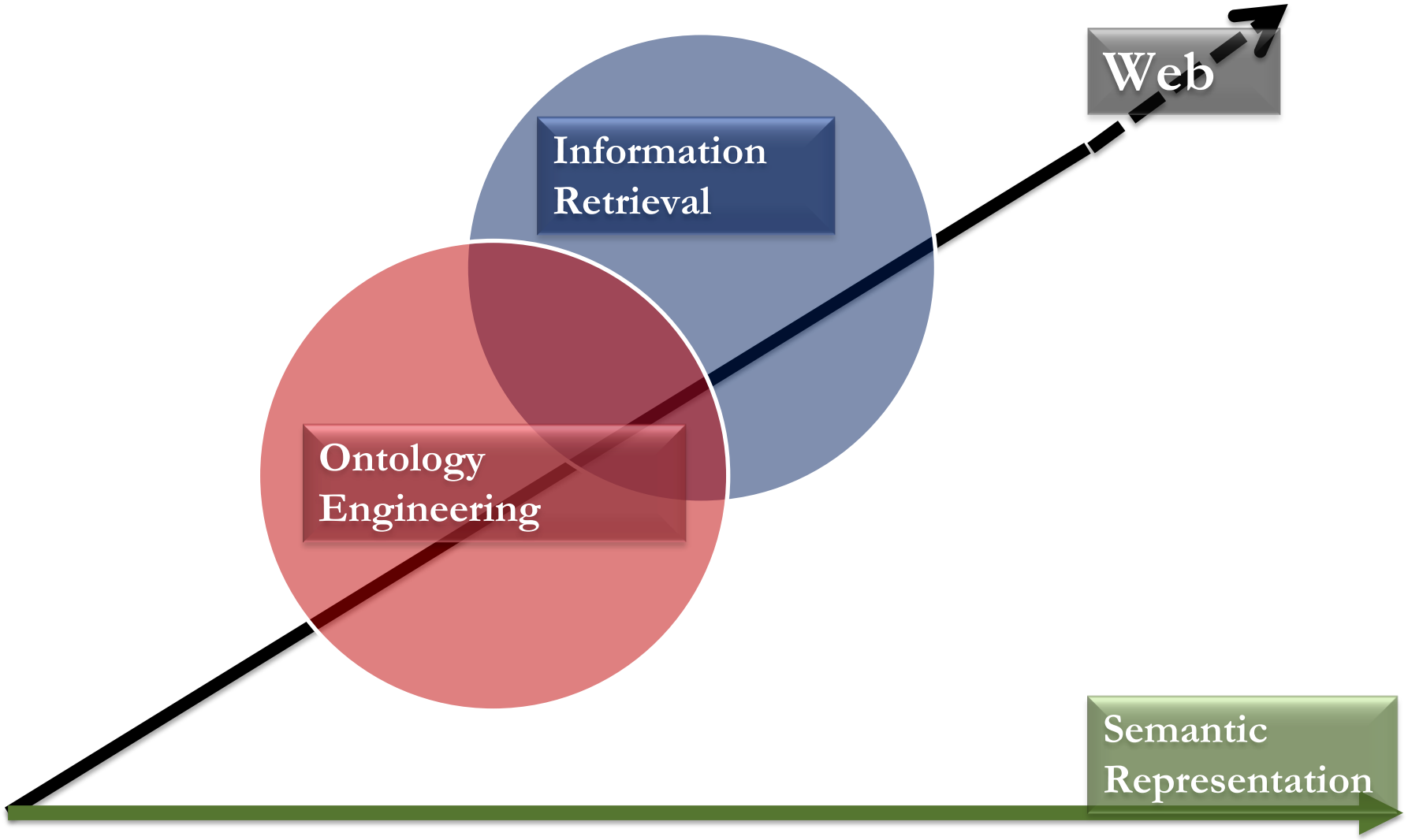
Evolution of the web
Existing large knowledge bases



Evolution of the web
Existing large knowledge bases



Correlated dimensions related to semantic BI



Evolution of Semantics: From Dictionaries To Ontologies

Evolution of Semantics

Web

Web 3.0 (2010)

Semantic Web (Web2.0)(2000)

Web 1.0 (1990)

Linked Open Data (LOD)

YAGO-NAGA

Dictionaries

Thesaurus

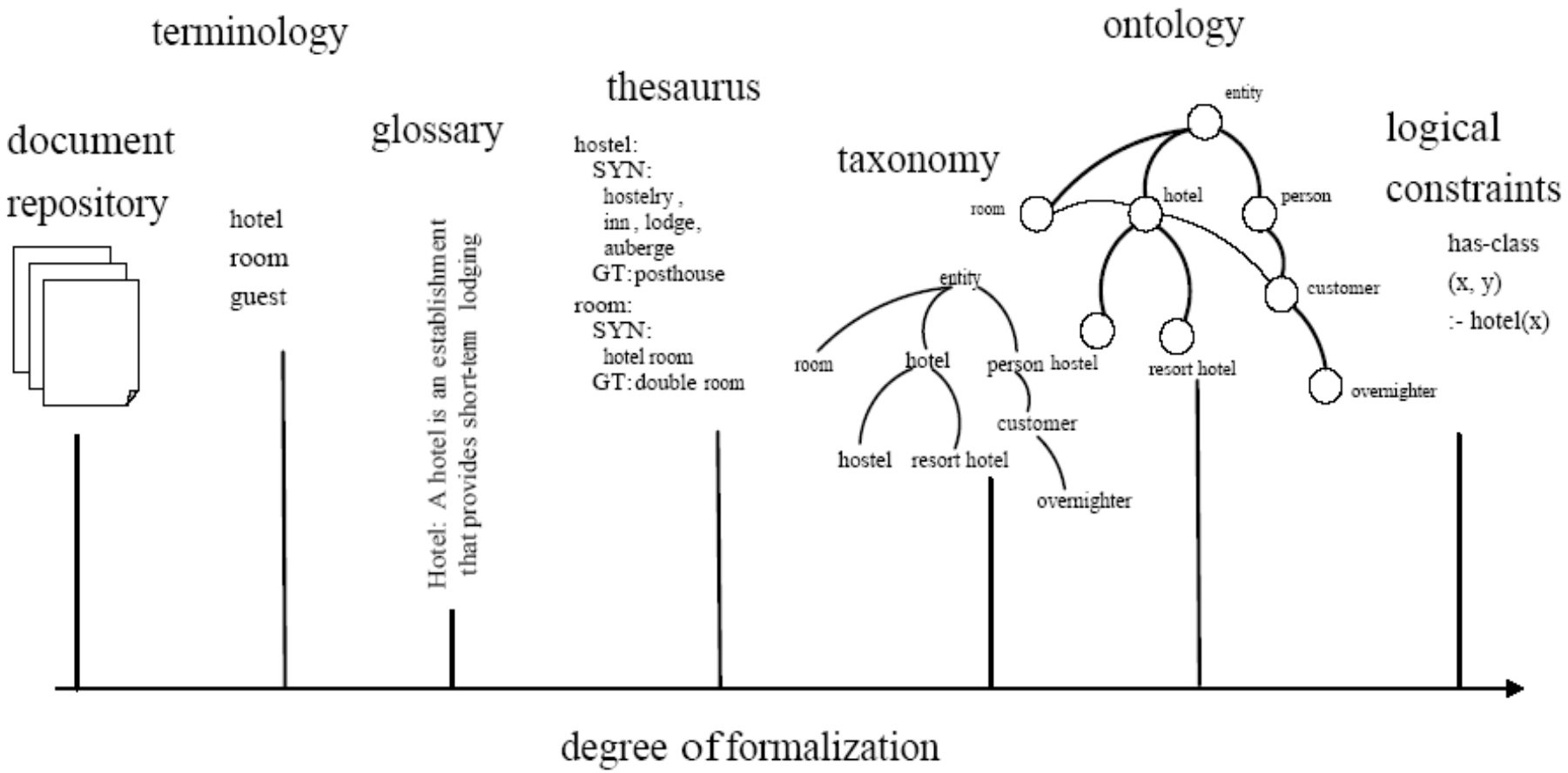
Controlled Vocabularies

Taxonomy

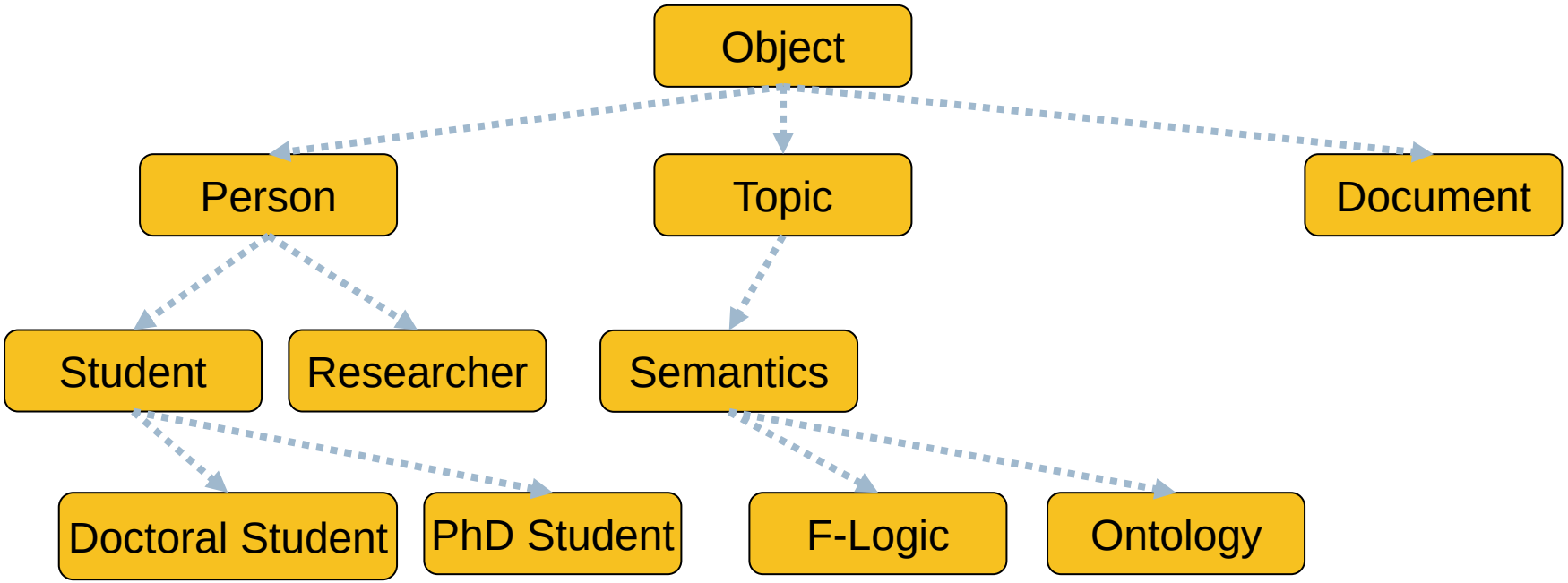
Ontologies

Semantics

Evolution of Semantics:
Levels of Knowledge Representation

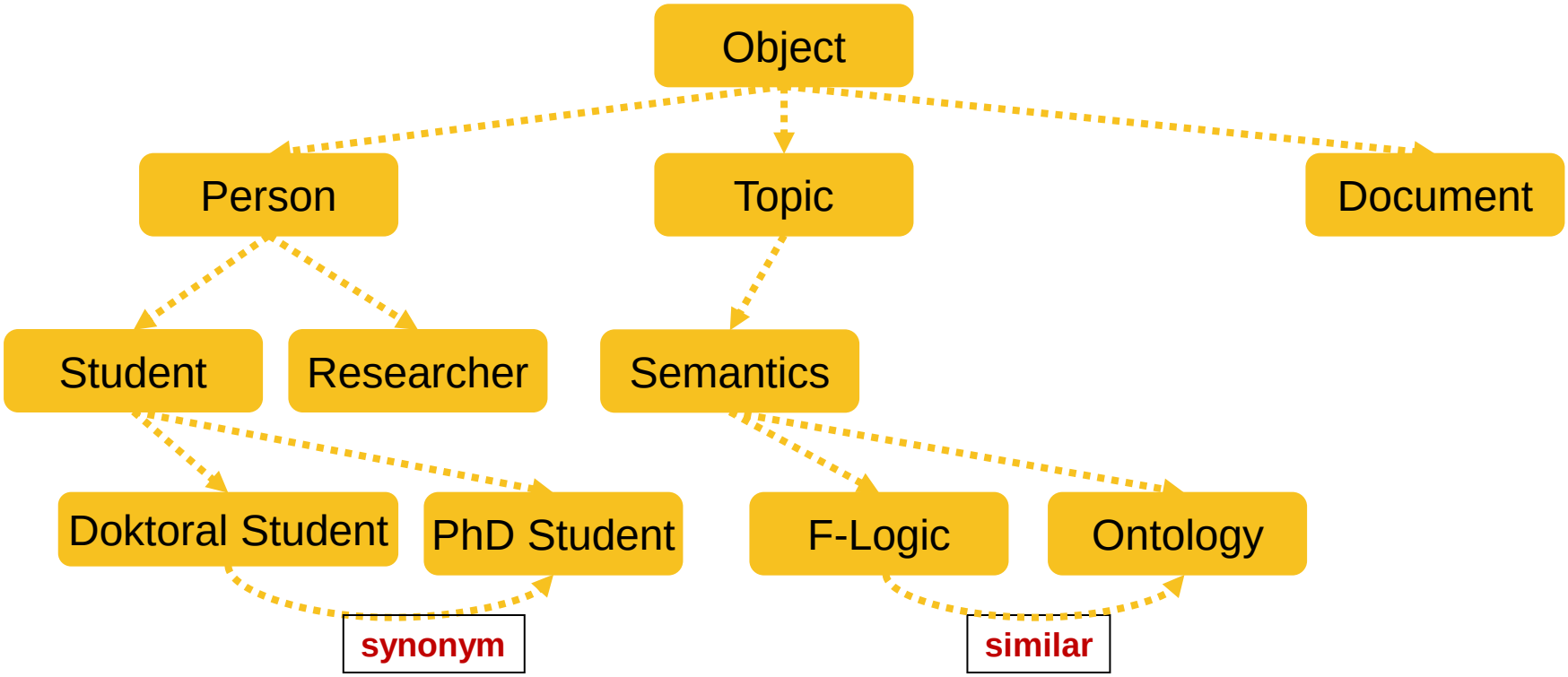


Evolution of Semantics:
Levels of Knowledge Representation



A Taxonomy

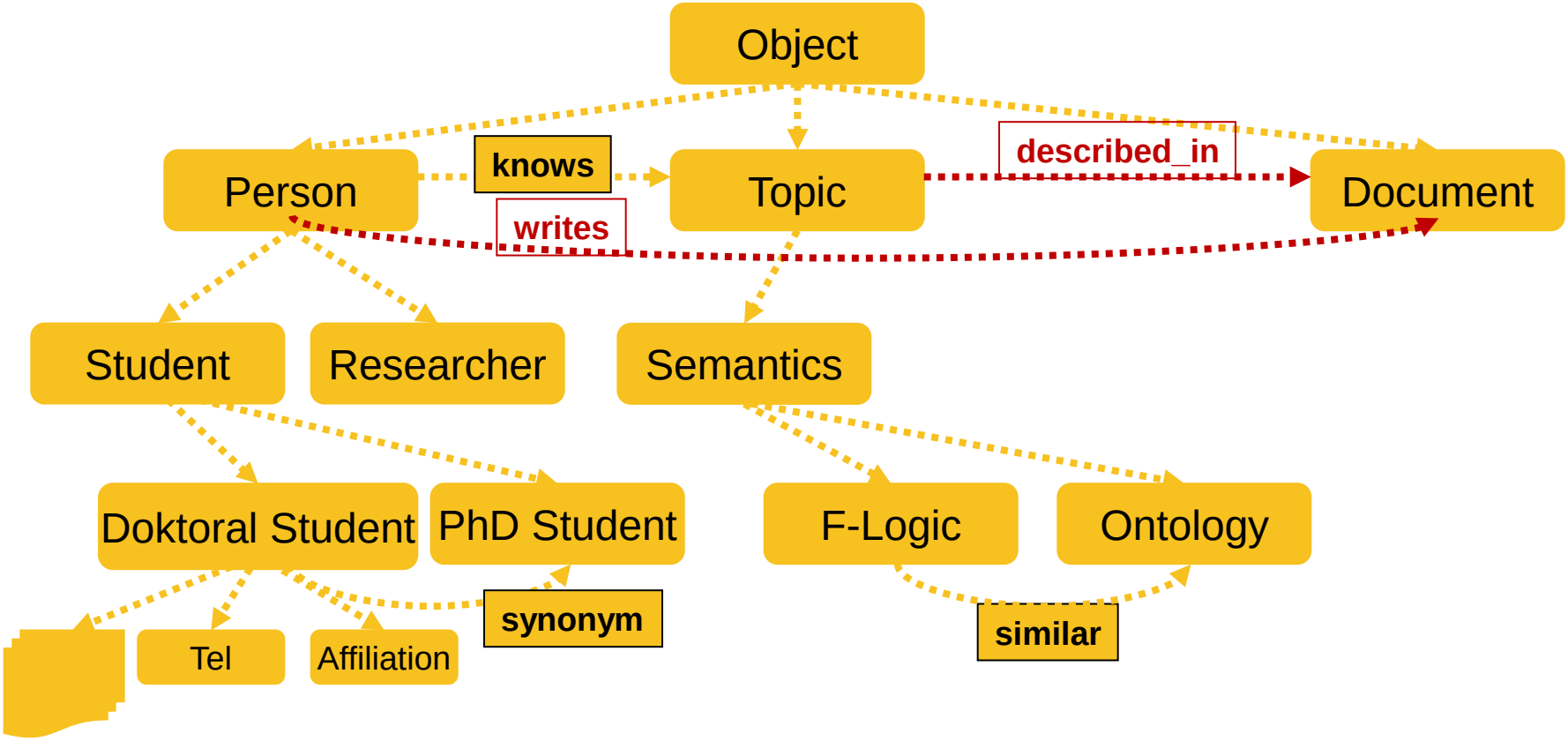
Evolution of Semantics:
Levels of Knowledge Representation



Thesaurus

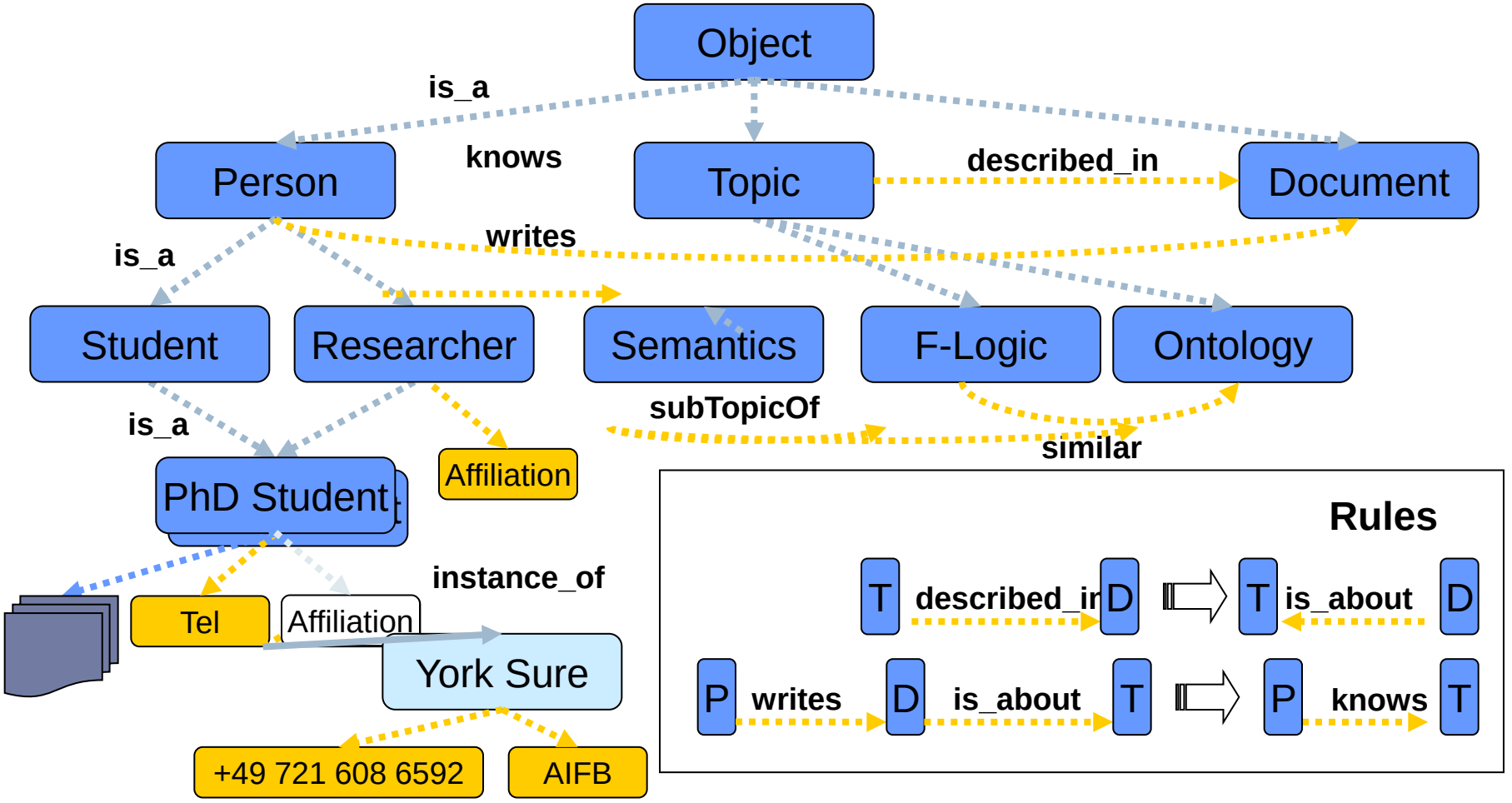
Graph with primitives, 2 fixed relationships (similar, synonym

Evolution of Semantics:
Levels of Knowledge Representation

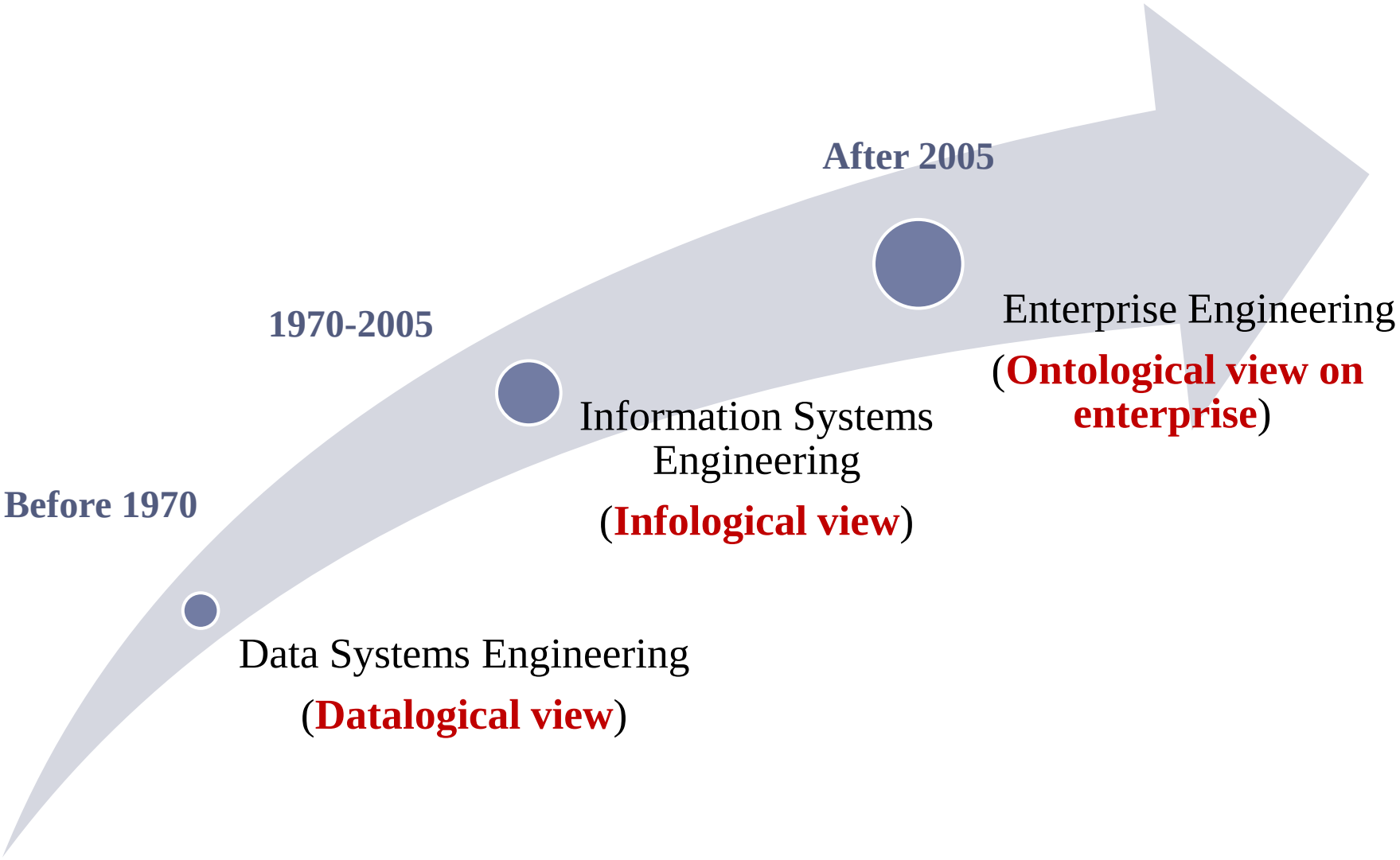


Topic Map

Evolution of Semantics:
Levels of Knowledge Representation



Evolution of Semantics:
Impact on business domain

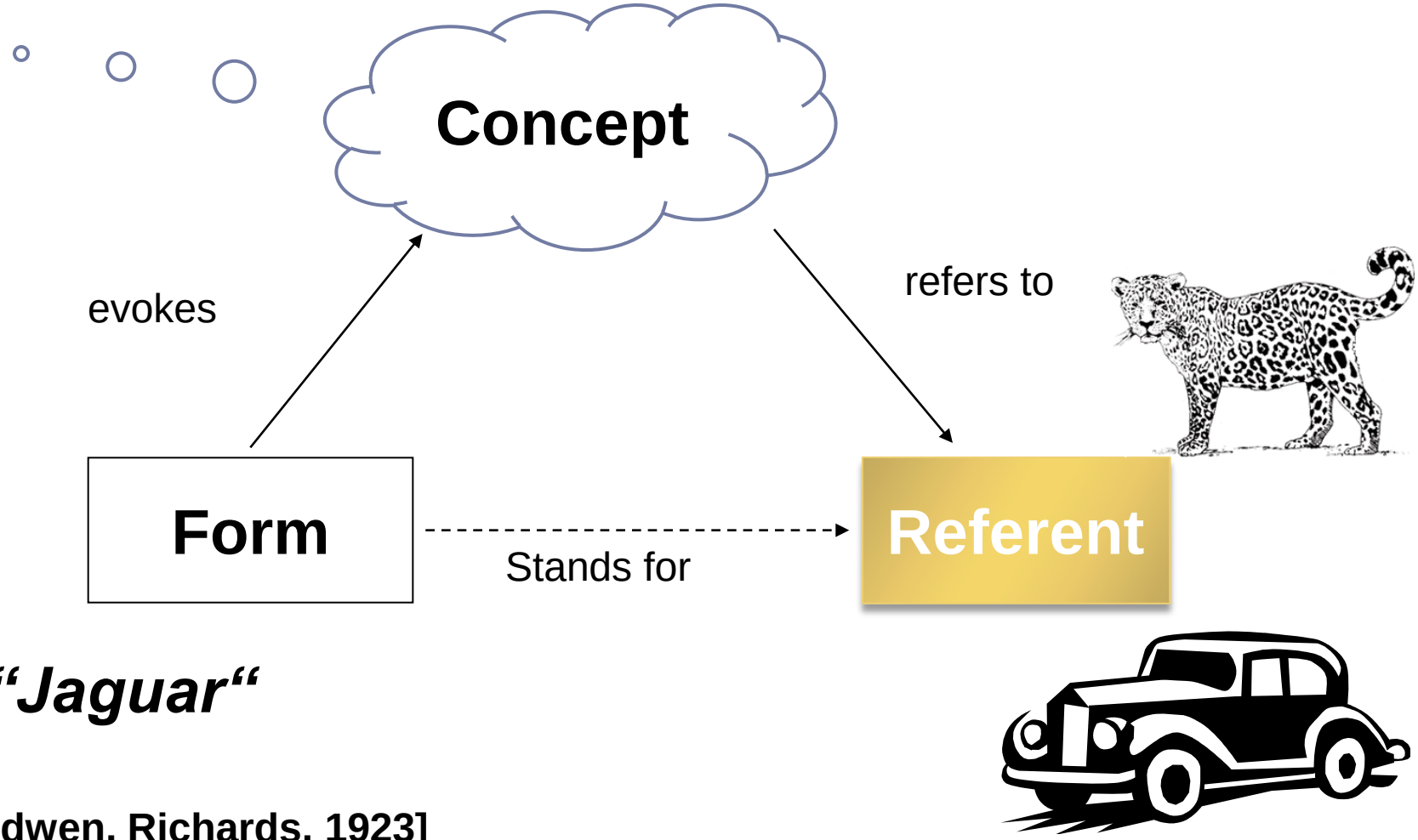


Evolution of Semantics: **Ontology definition**

- Origin:
  - *A branch of philosophy that investigates and explains the nature and essential properties and relations of all beings, or the principles and causes of being.*

- Computer Science (Artificial Intelligence):
  - *An ontology is a formal and explicit specification of a shared conceptualization (Gruber, 93) among a community of people (and agents) of a common area of interest.*

Evolution of Semantics:
Ontology: communication principle



[Odwen, Richards, 1923]

Evolution of Semantics:
Ontology structure

- Concept : *class of objects referred by a set of terms (synonyms)*
  - Customer
  - Product
  - Marketing strategy
  - Person
  - Invoice
- Taxonomic (hyponymy) relations: is-a relation
  - Customer is a person
- Meronymy relations: part-of relation
  - An invoice ***includes*** Marketing Strategies

Evolution of Semantics:
Ontology structure

- Non taxonomic relations: associative properties between concepts
  - A customer ***subscribes to*** products
  - An invoice ***belongs to a*** customer
  - An invoice ***has*** products
- Axioms: consist in
  - defining the meaning of concepts;
  - setting restrictions on attribute value
  - verifying the validity of specified knowledge

Evolution of Semantics: **Ontology structure**

- Rules: used to infer additional statements
  - *The discount for a customer buying a product is 7.5 percent if the customer is premium and the product is luxury.*

```
<Implies>
 <head>
  <Atom>
   <Rel>discount</Rel>
   <Var>customer</Var>
   <Var>product</Var>
   <Ind>7.5 percent</Ind>
  </Atom>
 </head>
 <body>
  <And>
   <Atom>
    <Rel>premium</Rel>
    <Var>customer</Var>
   </Atom>
   <Atom>
    <Rel>luxury</Rel>
    <Var>product</Var>
   </Atom>
  </And>
 </body>
</Implies>
```

Evolution of Semantics: **Ontology Types**

- Four typology in the literature:
  - Formalization degree: formal, informal and semi-formal
  - Granularity degree
  - Level of completeness
  - Domain of knowledge
    - Top-level Ontology (generic ontology)
    - Lexical Ontology: WorldNet
    - Domain Ontology
    - Ontology of tasks
    - Application Ontology

Evolution of Semantics:
Ontologies - Some Examples

- **General purpose ontologies:**
  - WordNet, http://www.cogsci.princeton.edu/~wn
  - EuroWordNet

- **Upper level ontologies:**
  - DOLCE
  - Upper-Cyc Ontology, http://www.cyc.com/cyc-2-1/index.html
  - IEEE Standard Upper Ontology, http://suo.ieee.org/

- **Domain and application-specific ontologies:**
  - RDF Site Summary RSS, http://groups.yahoo.com/group/rss-dev/files/schema.rdf
  - UMLS, http://www.nlm.nih.gov/research/umls/
  - AIFB Web Page Ontology, http://ontobroker.semanticweb.org/ontos/aifb.html
  - Web-KB Ontology, http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/
  - Dublin Core, http://dublincore.org/
  - Gene Ontology: http://www.geneontology.org/
  - Mesh Ontology: http://bike.snu.ac.kr/?q=node/207

Evolution of Semantics:
Enterprise Ontology

Enterprise Ontology:

a formal and explicit specification of a shared conceptualization among a community of people (managers, developers, employees and users)

Evolution of Semantics:
Enterprise Ontology: *CoProE Ontology*

- ***CoProE Ontology:***

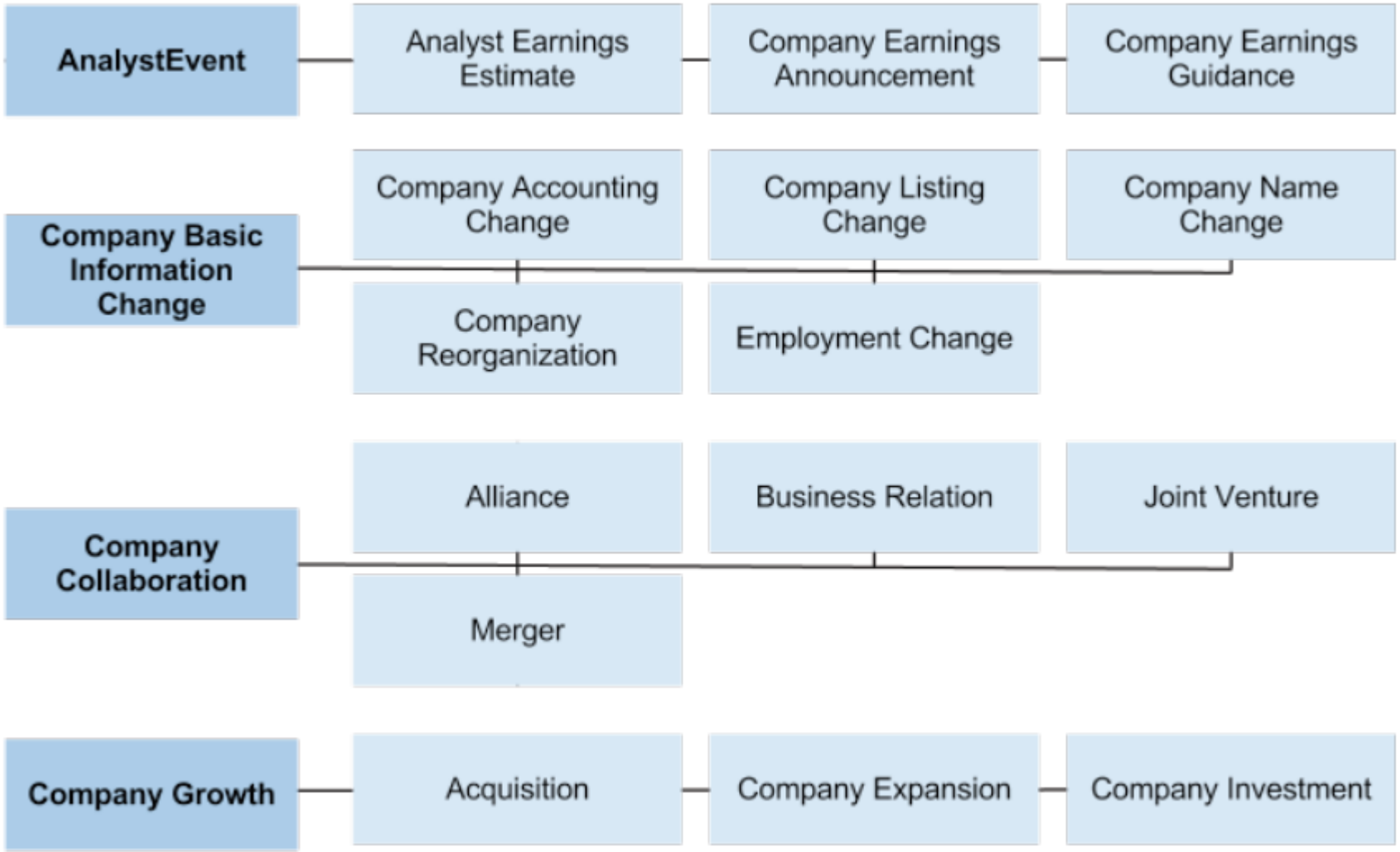
a *newsEvents ontology* [Lösch, 2009] and *United Nations Standard Products and Services Code* (UNSPSC) [1] classification of products and segments of industries.

- "Towards e-leadership: Higher profitability through innovative management and leadership systems" project

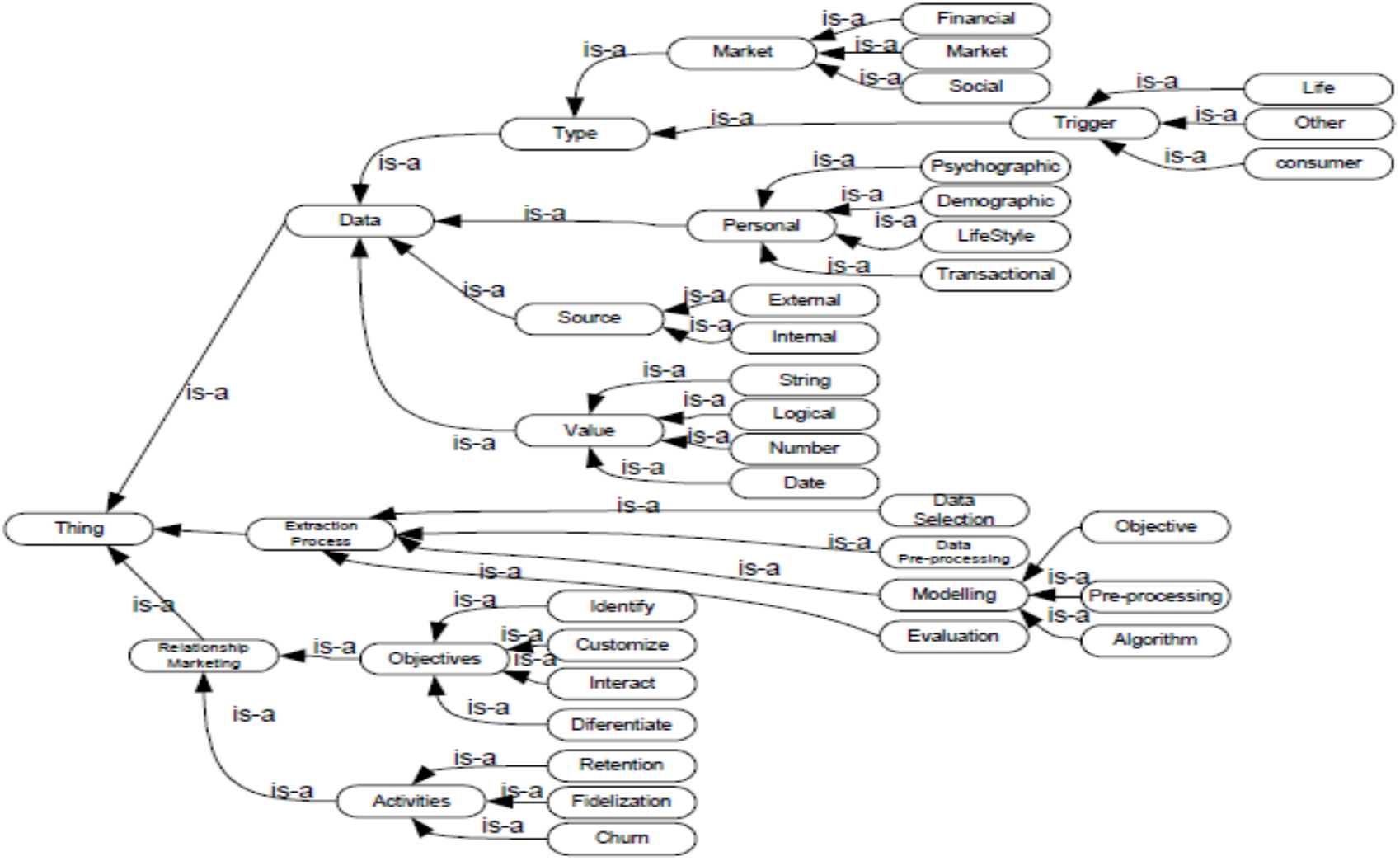
Lösch, U., Nikitina, N.: The newsEvents Ontology – An Ontology for Describing Business Events. In: 8th International Semantic Web Conference, 1st Workshop on Ontology Design Patterns, Washington DC, USA (2009)

1. <http://www.unspsc.org/

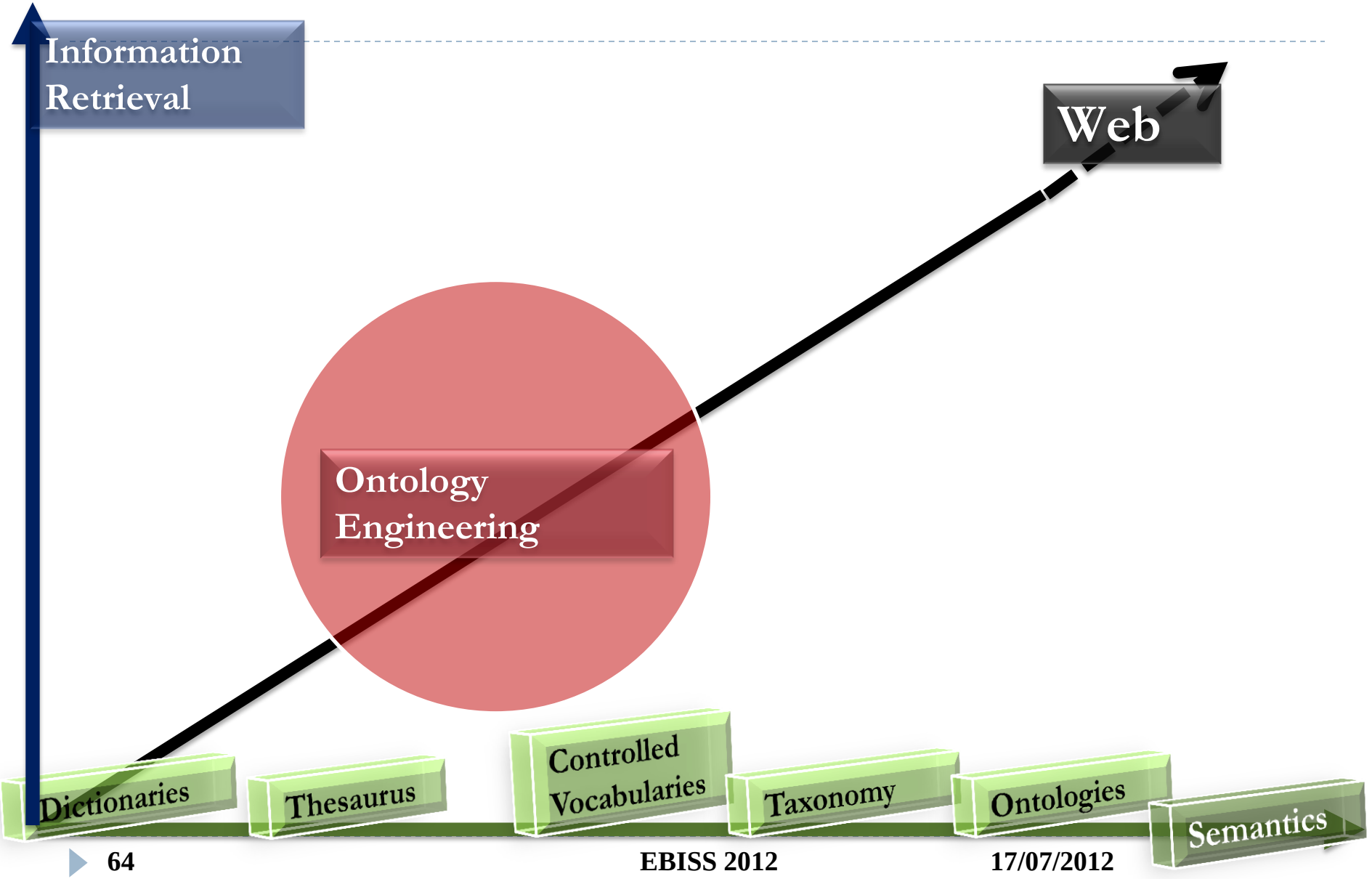
Evolution of Semantics:
Enterprise Ontology: *CoProE Ontology*



Evolution of Semantics:
Marketing ontology



Correlated dimensions related to semantic BI



New trends of search paradigm

Semantic web search classification in Web 2.0

Ontology search engines

Ontology Meta-search :

OntoSearch [Y. Zhang et al., 2004]

Swangler [T. Fini et al., 2005]

Crawler-based ontology search:

Ontokhoj [C. Patel et al., 2003]

Swoogle [T. Finin et al., 2005]

Semantic search engines

Contextual semantic search

QuizRDF [J. Davies et al., 2005],

Corese [O. Corby],

Infofox [B. Sigrist et al., 2003],

SHOE [B. Aleman-Meza et al., 2003],

DOSE [D. Bonino 2003], SERSE [V. Tamma, 2004]

Evolutive semantic search

W3C Semantic Search [R. Guha et al., 2005]

Semantic association discovery

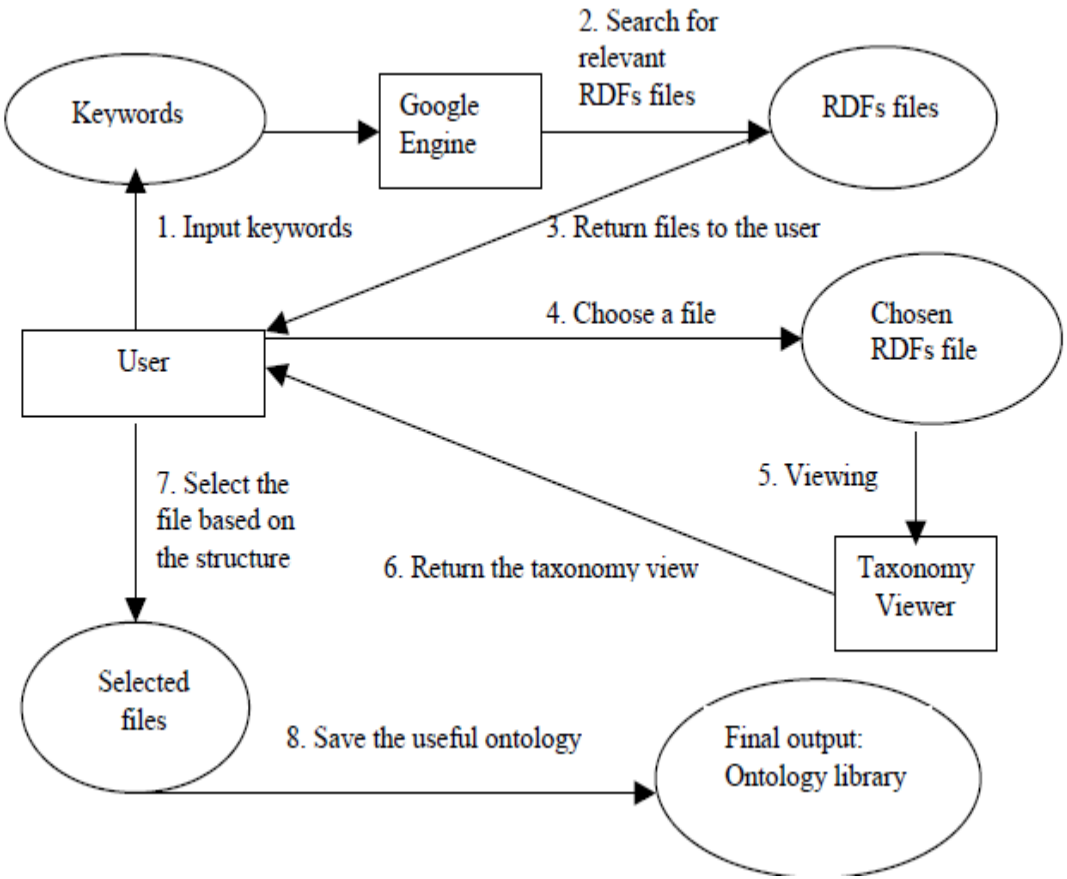
SemDis [C. Rocha et al., 2004]

Ontology Search Engines

Ontology Meta-search :
OntoSearch

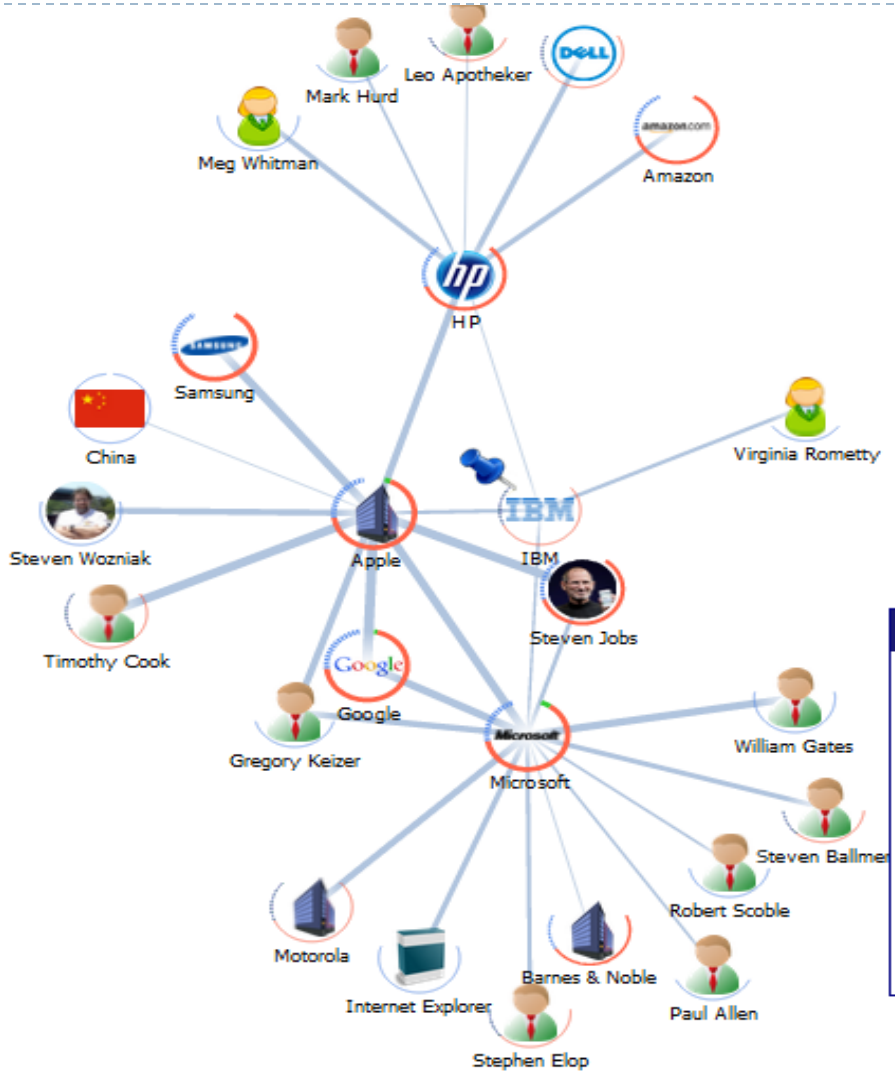
Swangler

- Use of conventional search engines
- Provide specific type of files (RSS, RDF, OWl)
- Search by name or options like “filetype”



Ontology Search Engines
OntoSearch

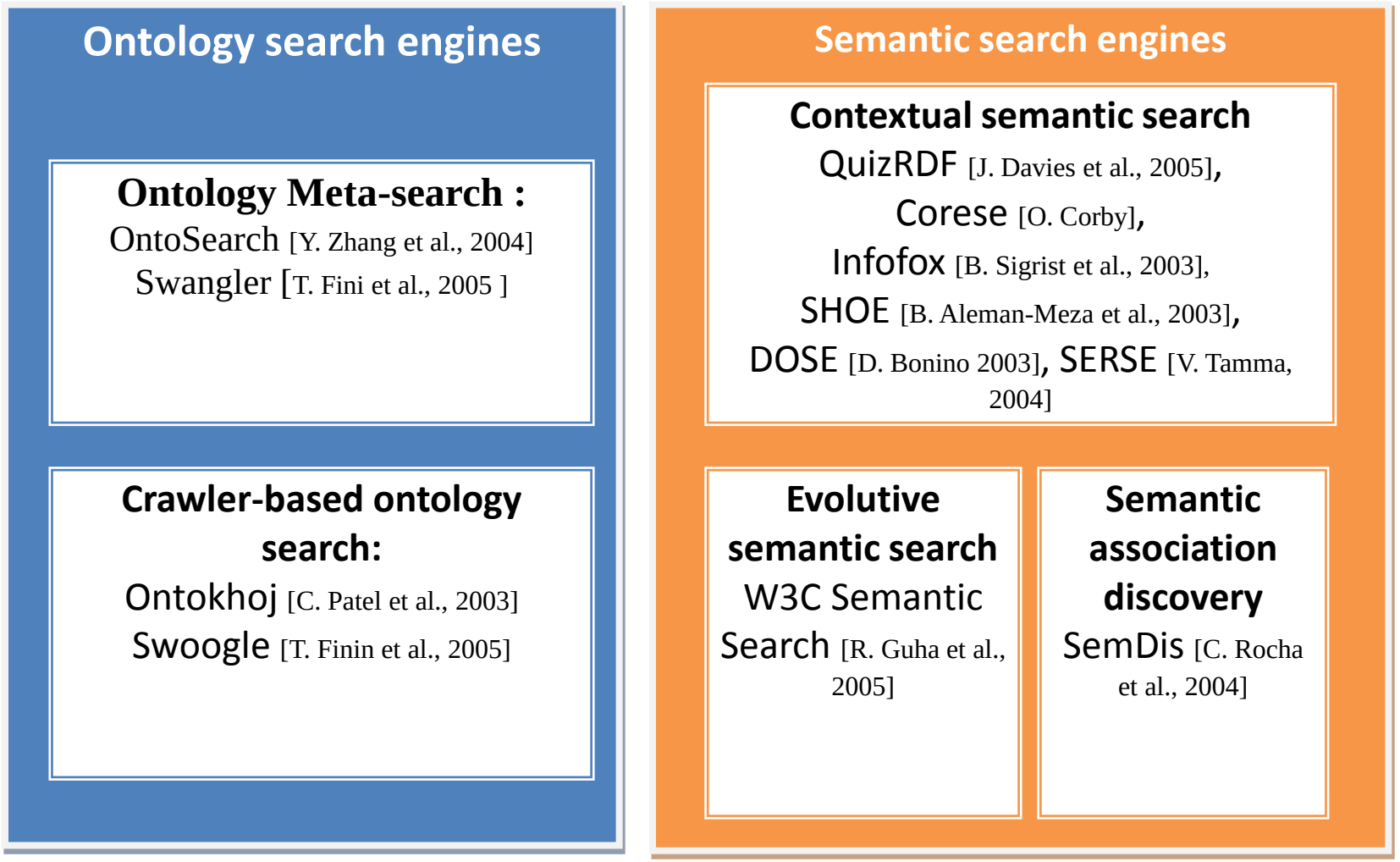
Ontology Search Engines
OntoSearch



William Gates (350 articles)

Robin Gibb, Bee Gees Co-Founder With His...
How Will Zuckerberg Spend His Millions? - Newsweek
Robin Gibb, Bee Gees Co-Founder With His...
16-year-old ping pong phenom to star for U.S.
A debate: Should you jump in on Facebook debut...
Should you buy Facebook? Pitting the genius of...
A debate: Should you jump in on Facebook debut...
O'Brien: The Other Faces of Facebook - San Jose...
Is Facebook IPO a sign of tech bubble 2.0...
Facebook narrowly avoids dip below starting price...

Semantic web search classification in Web 2.0



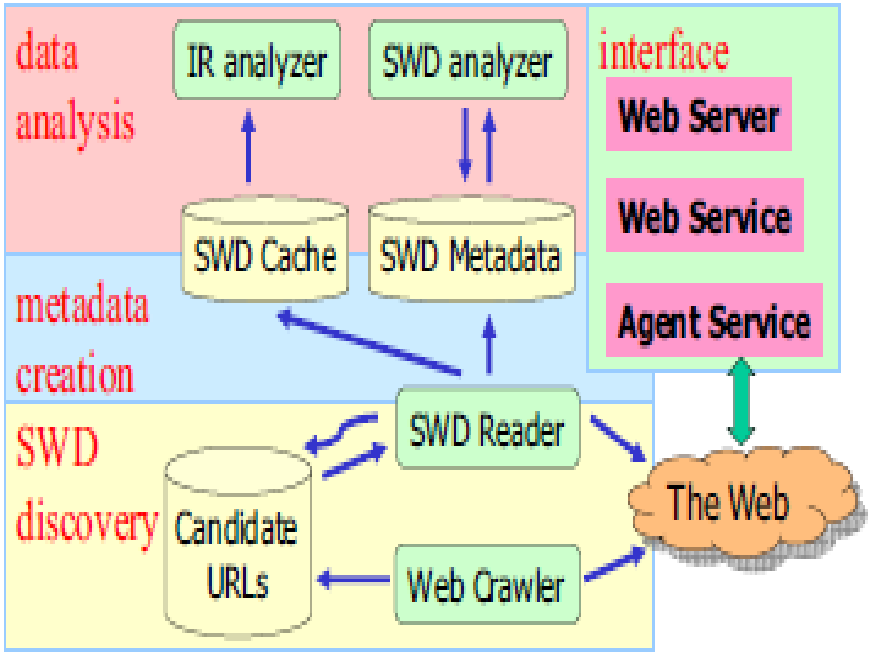
Ontology Search Engines

Crawler-based ontology search :

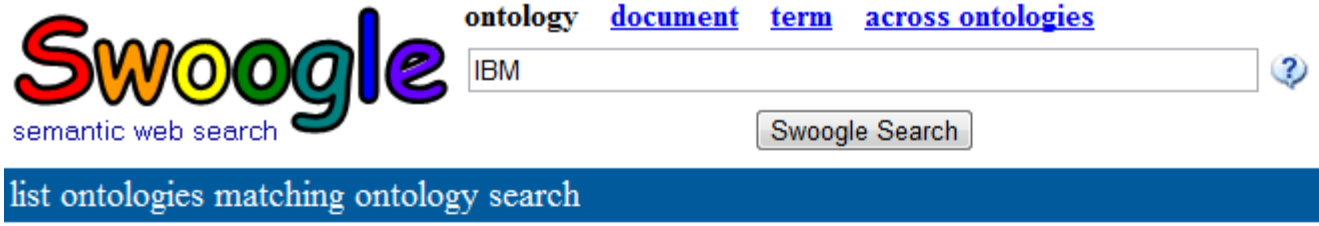
Ontokhoj

Swoogle

- Specific crawler for semantic web document



Ontology Search Engines



http://sweet.jpl.nasa.gov/ontology/data_center.owl
[DEF] _Station, I, **IBM**, **IBM**_Research, IFREMER, IFREMER_Laboratoire_Ecologie_halieutique, IGBP, IGN, IGNS, IGNS
SemanticWebDocument, RDFXML, 2005-04-19, 107K, ontoRatio(1.00), metadata, cached

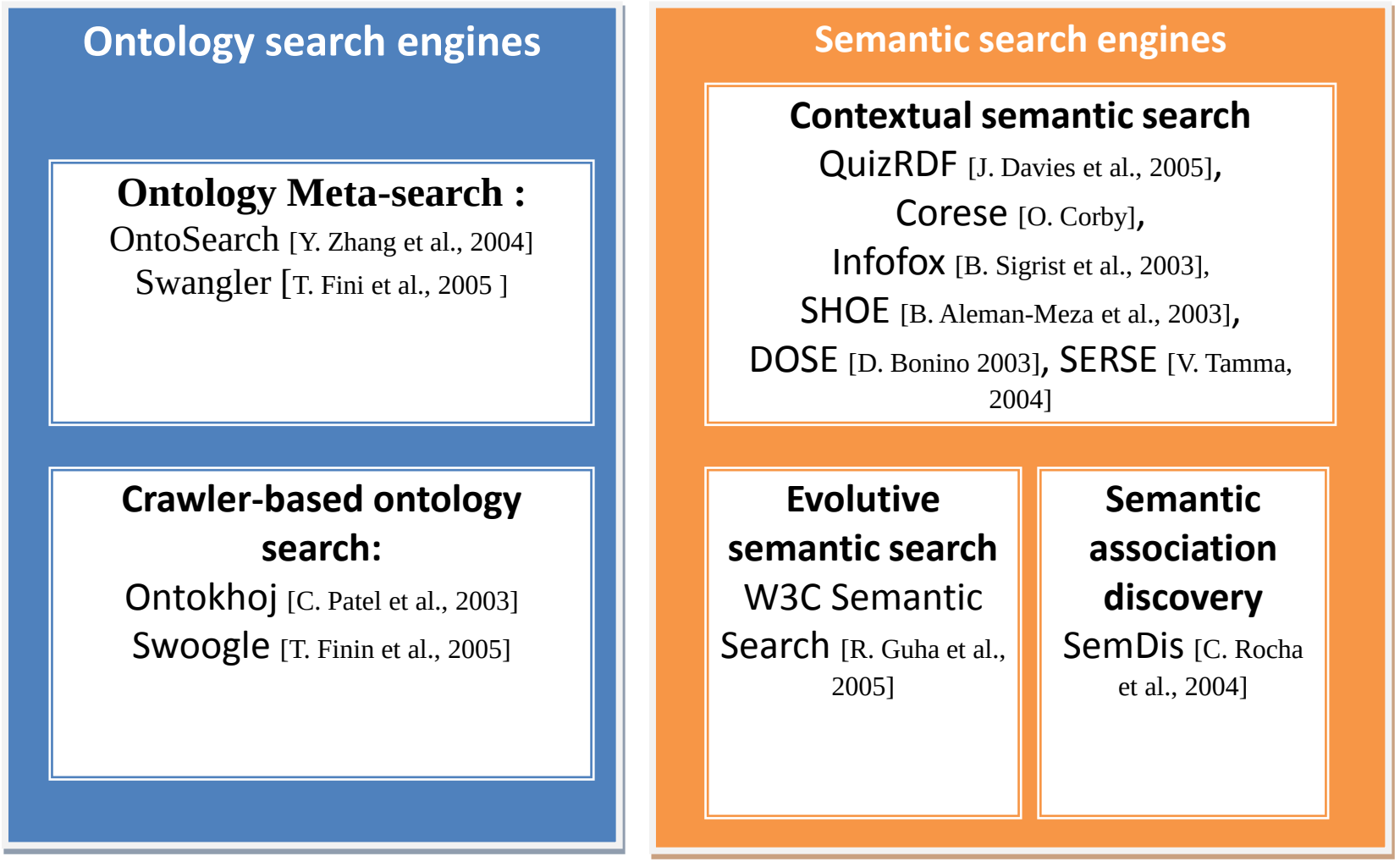
http://what.csc.villanova.edu/twiki/pub/Main/OWLFileInformation/ComputingOntology-George2.rdf-xml.owl
[DEF] _at_the_National_Physical_Laboratory, Harvard, Harvard_Machines, Harvard_Mark_1_operational_at_**IBM**_Endicot
SemanticWebDocument, RDFXML, 2008-06-23, 631K, ontoRatio(0.99), metadata, cached

http://www.pim.uni-essen.de/mitarbeiter/pimyial/Kompetenzontologie.rdf
[DEF] , Holzgewerbe, Holzgewerbe_ohne_Moebel, **IBM**, **IBM**_Betriebssystem_Kompetenz, IT, IT_Architektur_Kompetenz
SemanticWebDocument, RDFXML, 2005-02-22, 99K, ontoRatio(0.84), metadata, cached

http://monet.nag.co.uk/cocoon/monet/publicdocs/ontologies/hardware.owl
[DEF] , Enterprise10000, Fire, Fire12K, Fire15K, Fujitsu, Graphics, Hitachi, **IBM**, K, Manufacturer, Memory, Origin
SemanticWebDocument, RDFXML, 2005-06-07, 5K, ontoRatio(1.00), metadata, cached

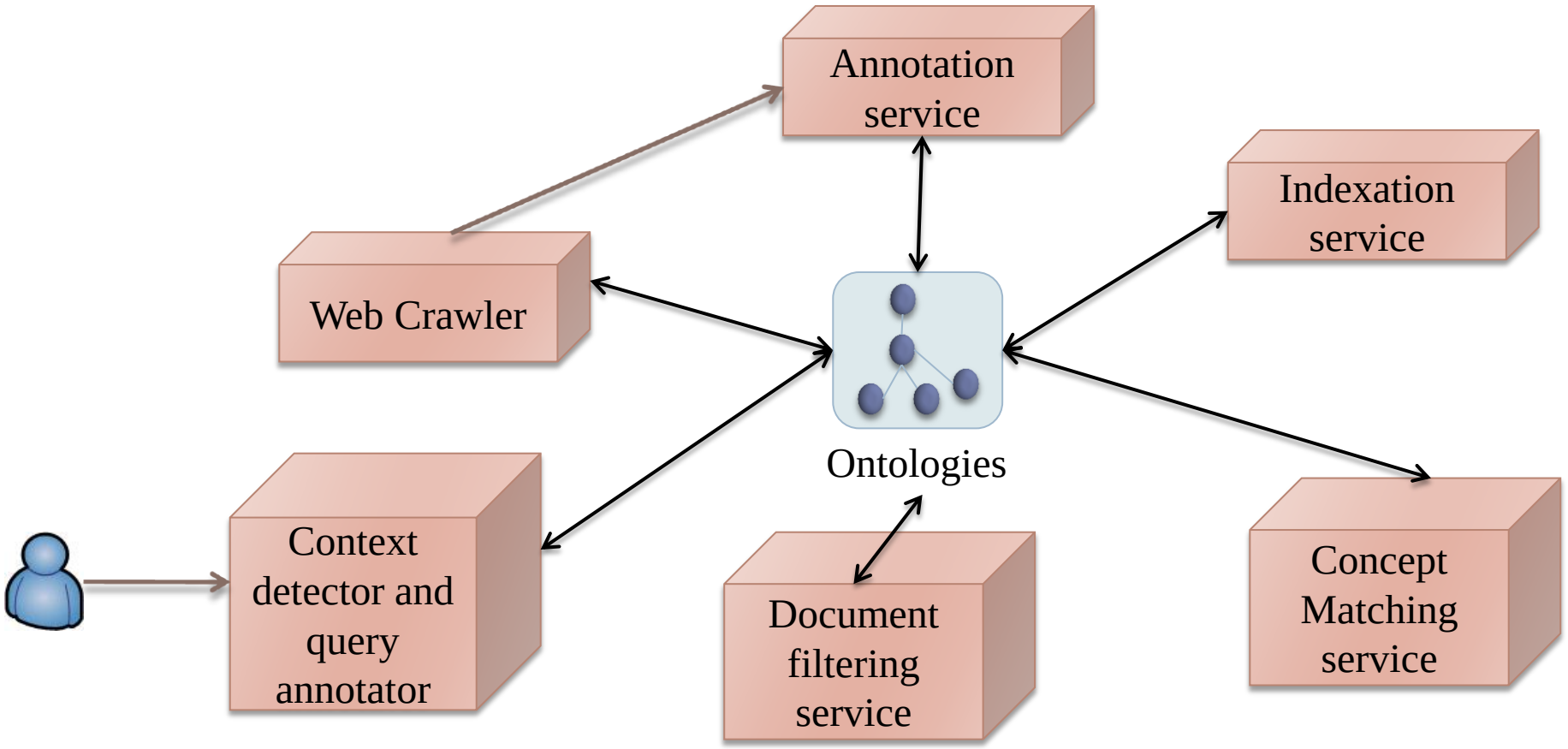
http://jastor.cvs.sourceforge.net/viewvc/*checkout*/jastor/com.ibm.adtech.jastor.test/ontologies/skin3.rdf
SemanticWebDocument, N3, 2008-11-17, 18K, ontoRatio(0.93), metadata, cached

Semantic web search classification in Web 2.0



Semantic search Engine

1. Contextual semantic search

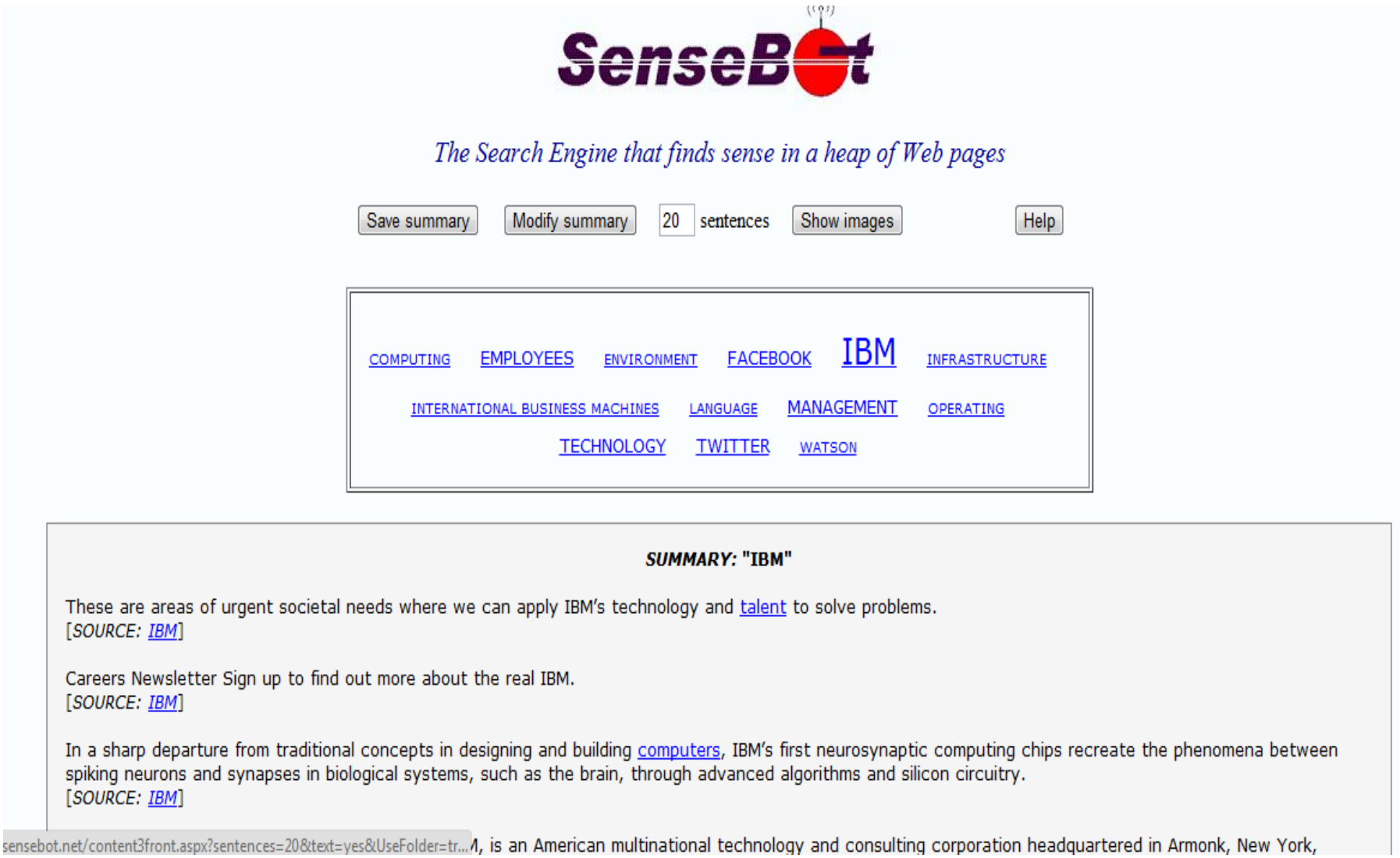


Semantic search Engine

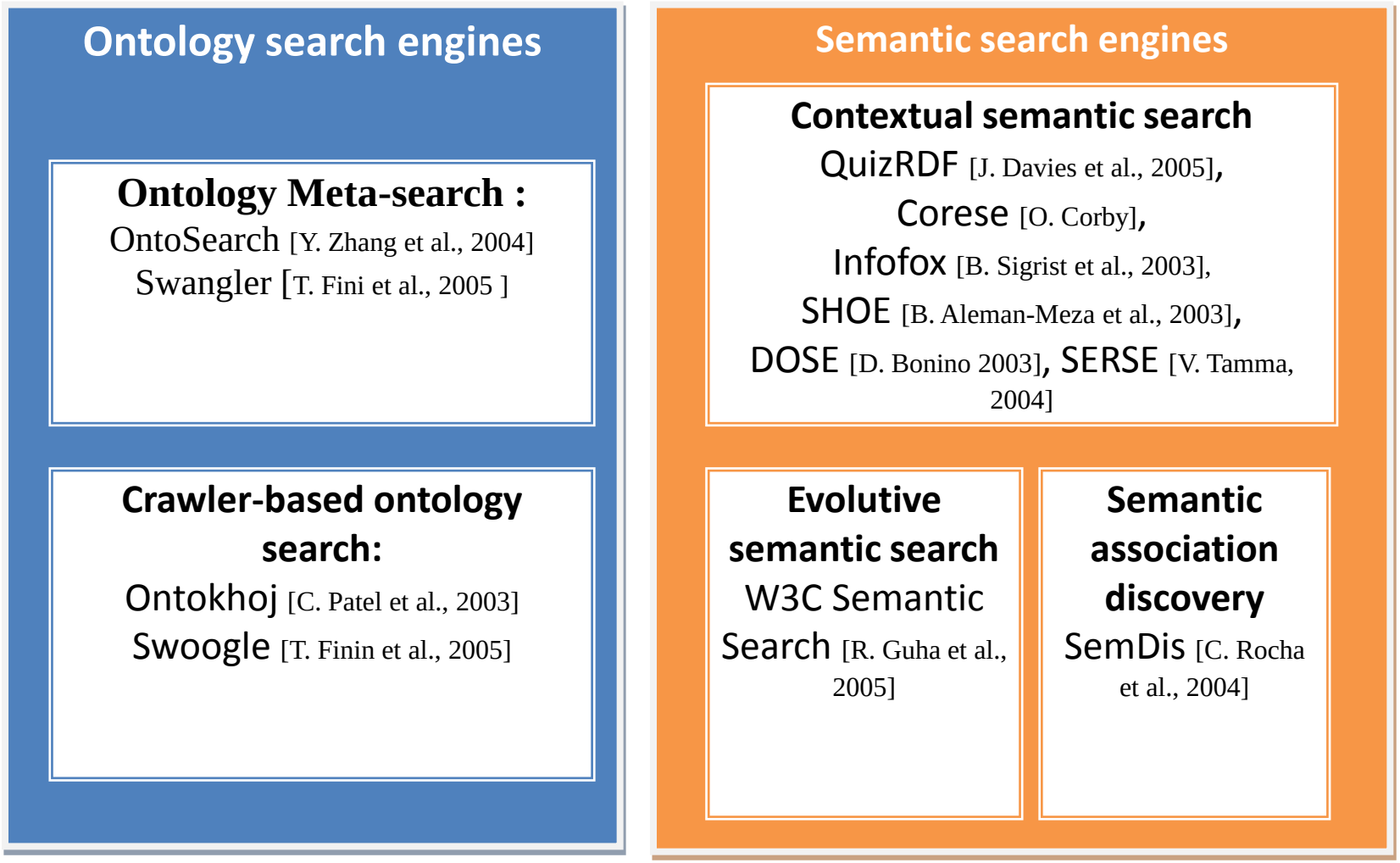
1. Contextual semantic search

Semantic search Engine

1. Contextual semantic search

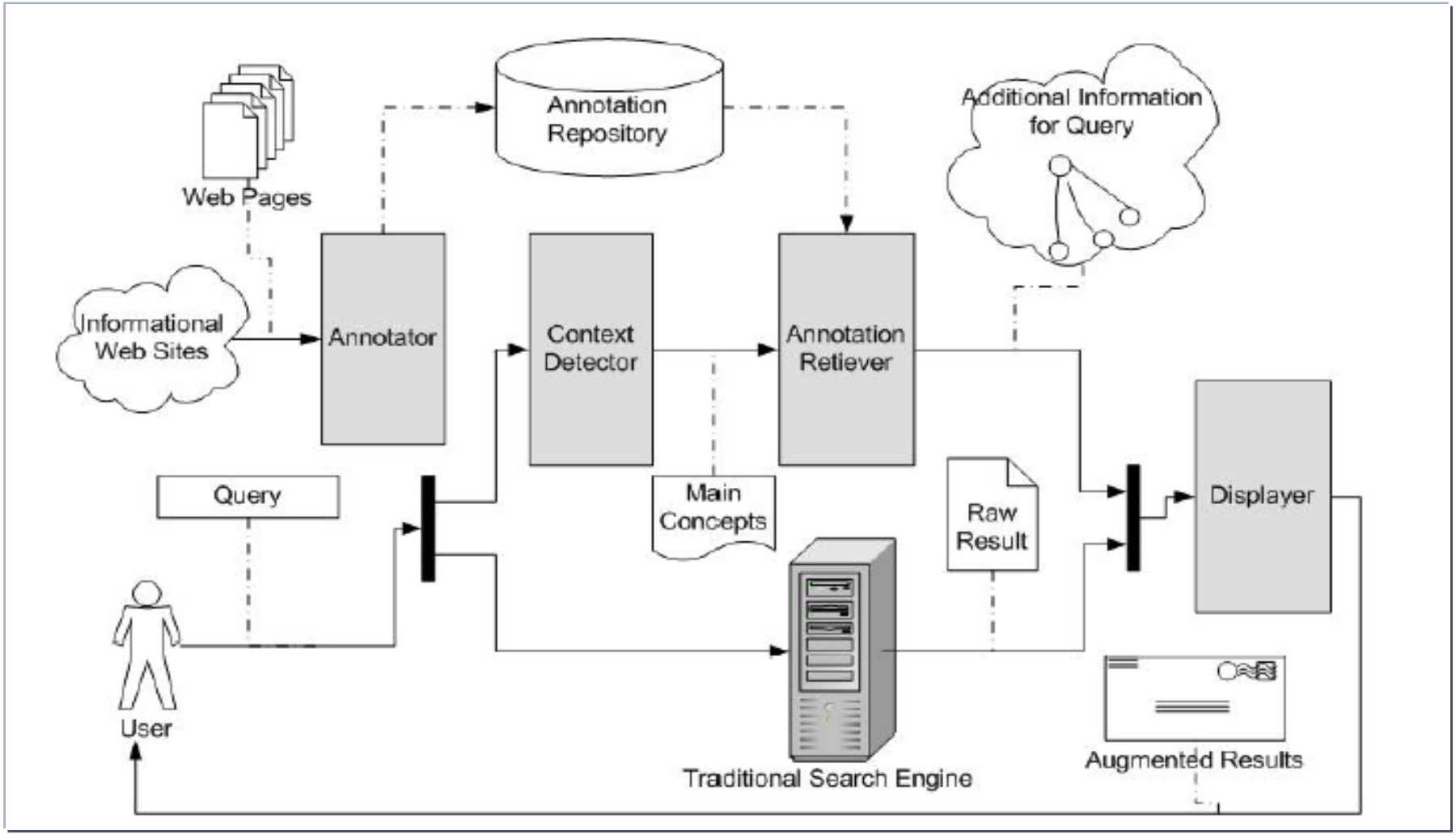


Semantic web search classification in Web 2.0



Semantic search Engine

2. Evolutive semantic search



Semantic web search classification in Web 2.0

Ontology search engines

Ontology Meta-search :

OntoSearch [Y. Zhang et al., 2004]

Swangler [T. Fini et al., 2005]

Crawler-based ontology search:

Ontokhoj [C. Patel et al., 2003]

Swoogle [T. Finin et al., 2005]

Semantic search engines

Contextual semantic search

QuizRDF [J. Davies et al., 2005],

Corese [O. Corby],

Infofox [B. Sigrist et al., 2003],

SHOE [B. Aleman-Meza et al., 2003],

DOSE [D. Bonino 2003], SERSE [V. Tamma, 2004]

Evolutive semantic search

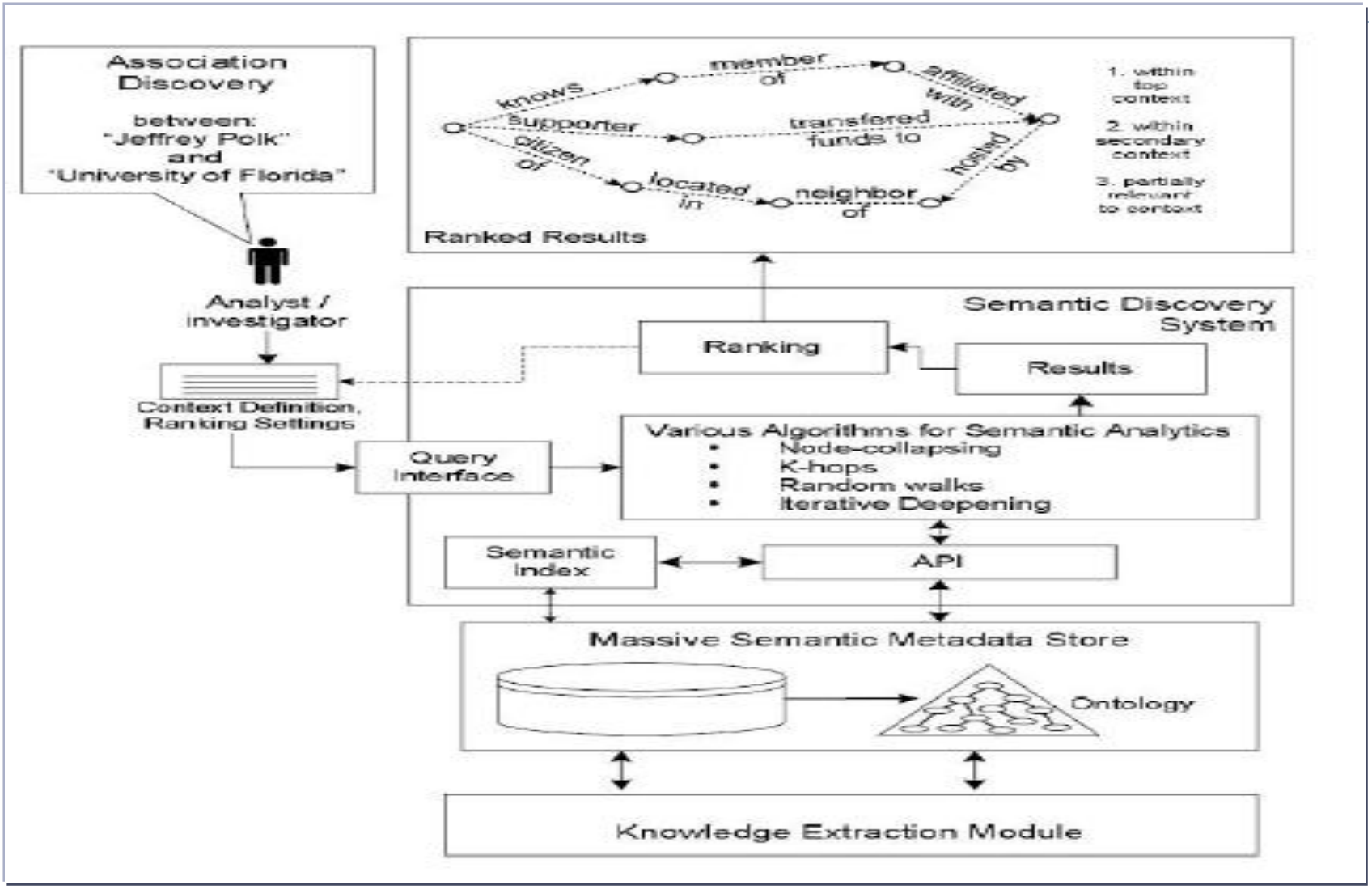
W3C Semantic Search [R. Guha et al., 2005]

Semantic association discovery

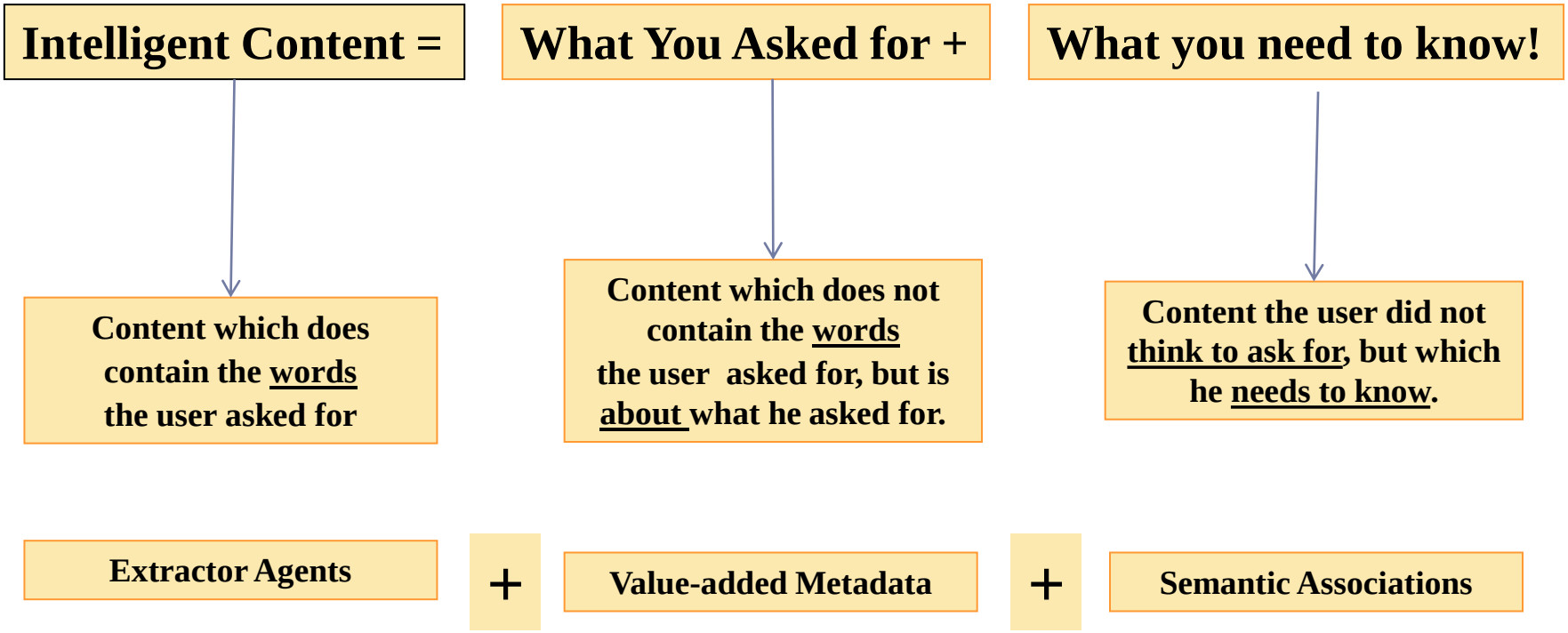
SemDis [C. Rocha et al., 2004]

Semantic search Engine

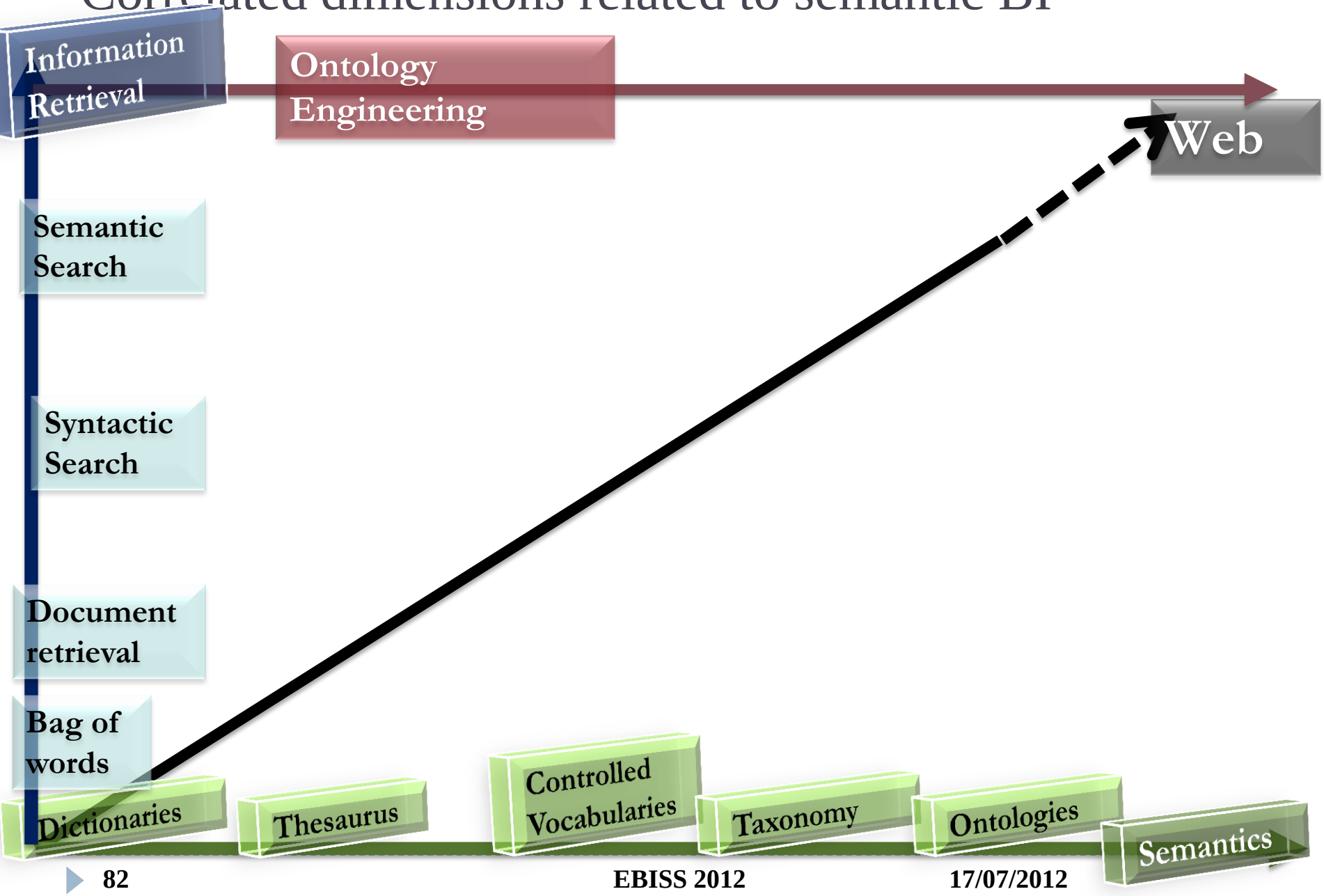
3. Association Discovery



Enterprise Search



Correlated dimensions related to semantic BI

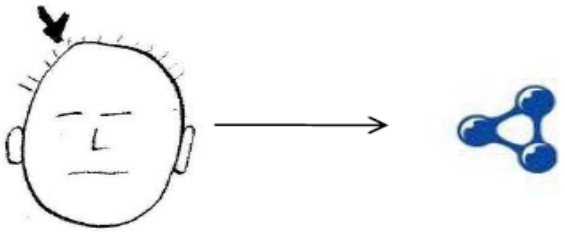


Progress in Ontology Engineering Research

Ontology engineering vs. learning

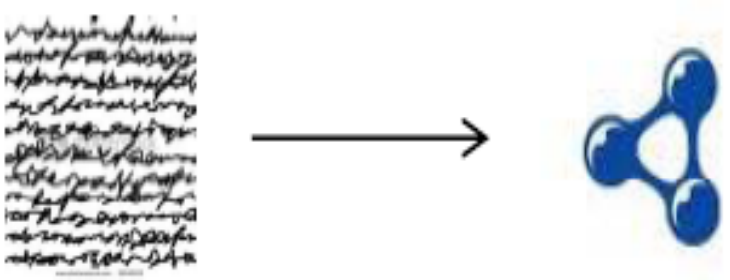
Ontology engineering

- Expert-driven, small-scale
- Knowledge and concept identification
- Preliminary informal representation
- Formalization (RDF, OWL, etc)
- Evaluation and Maintenance

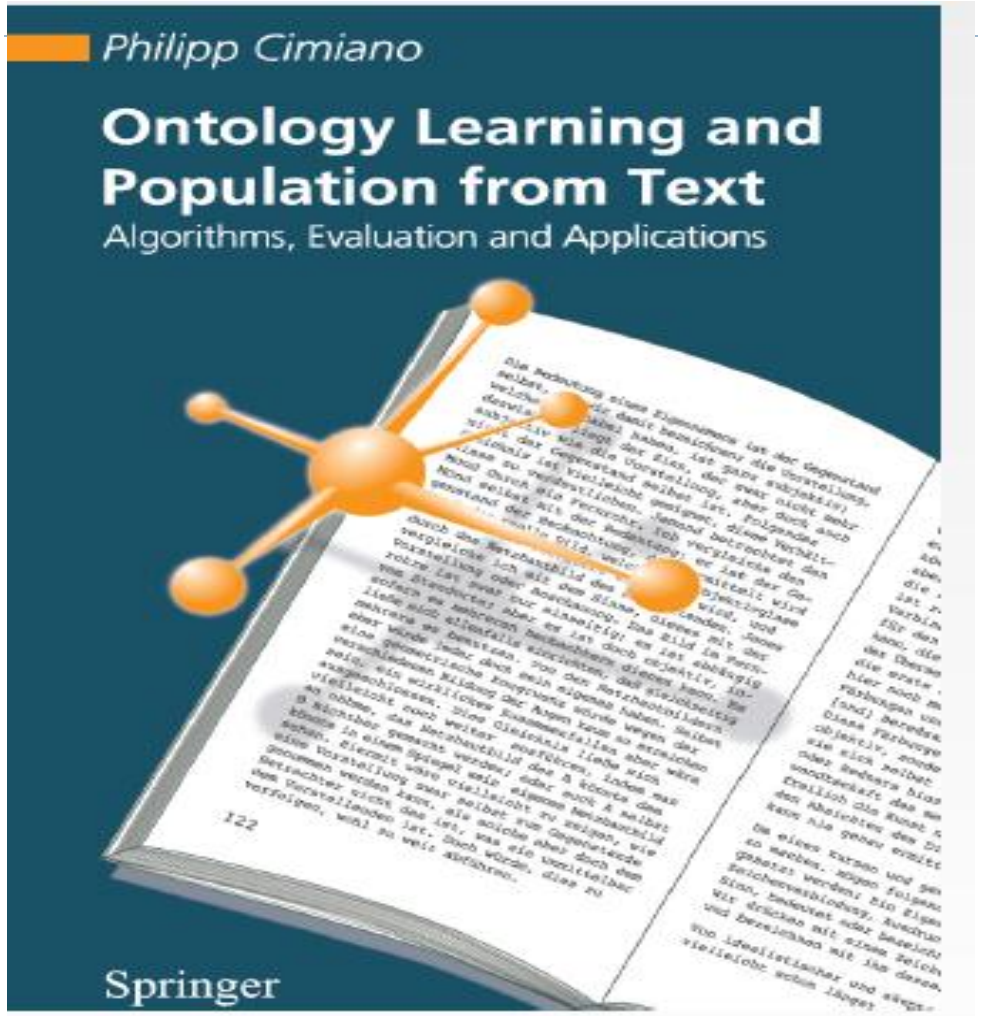


Ontology learning

- Data driven large scale
- Source selection
- Data exploration
- Concept and relation learning
- Evaluation and Updating

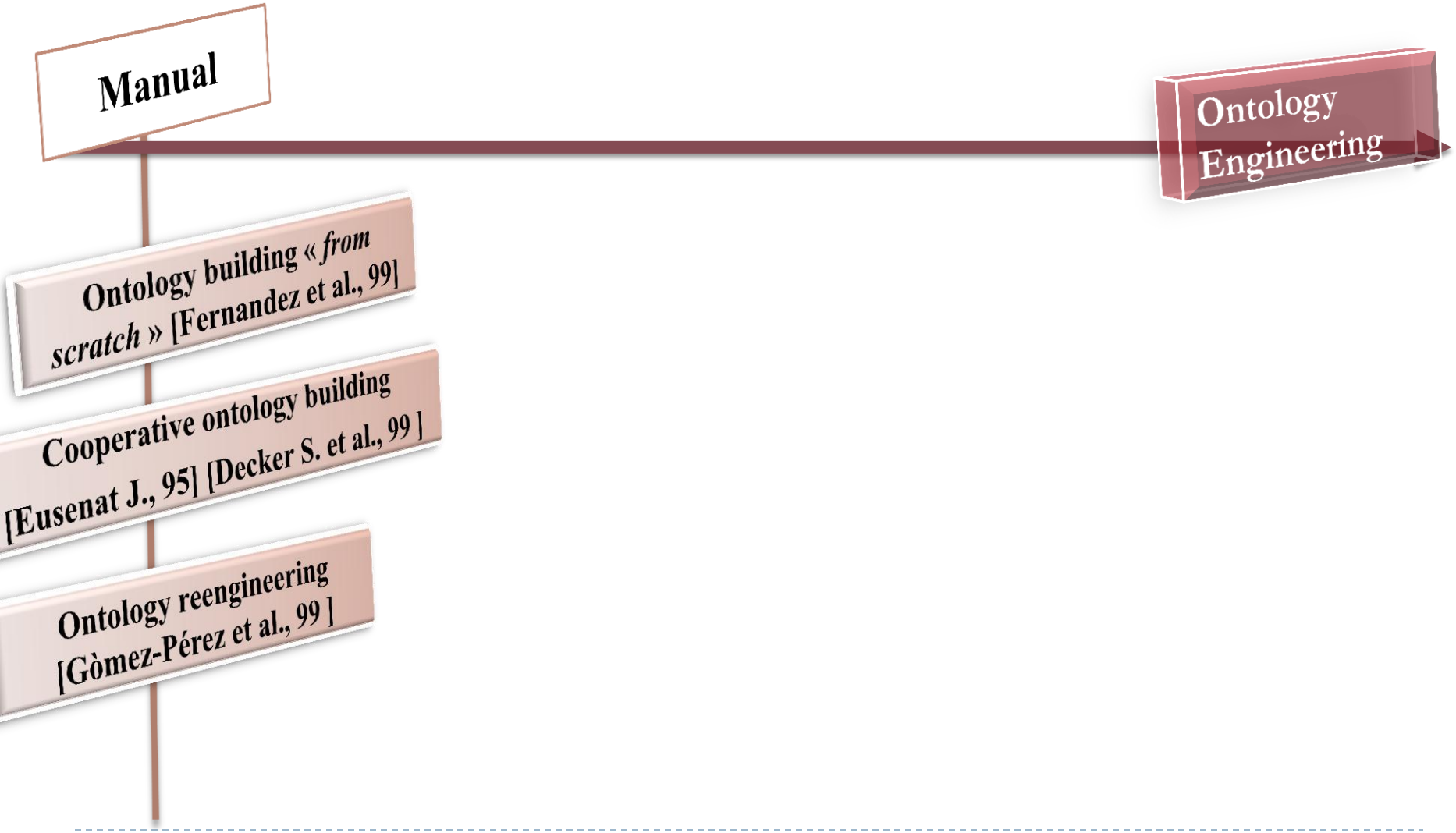


Ontology Learning

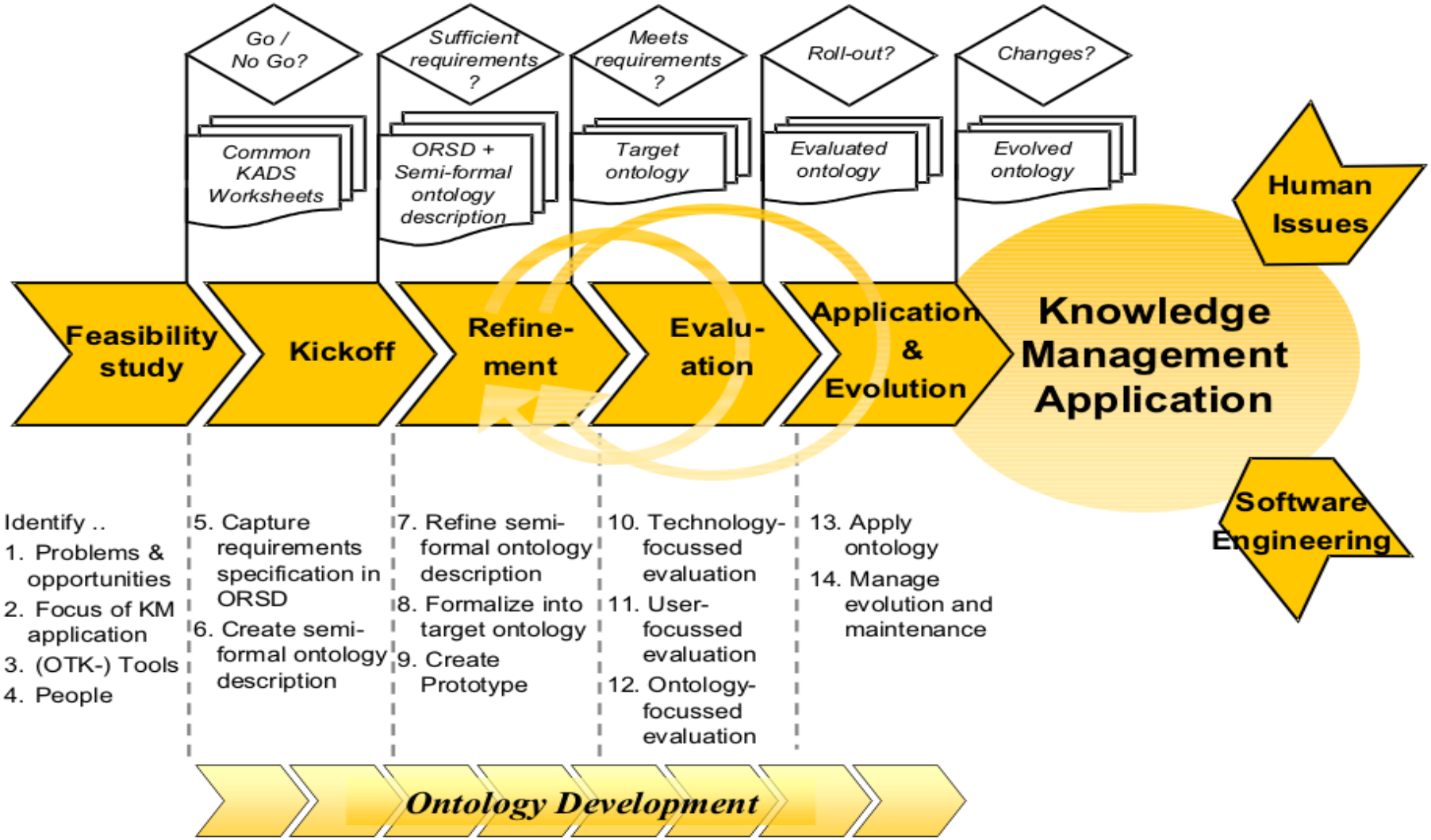


(Cimiano, 2006)

Ontology Learning
History



Manual Ontology engineering

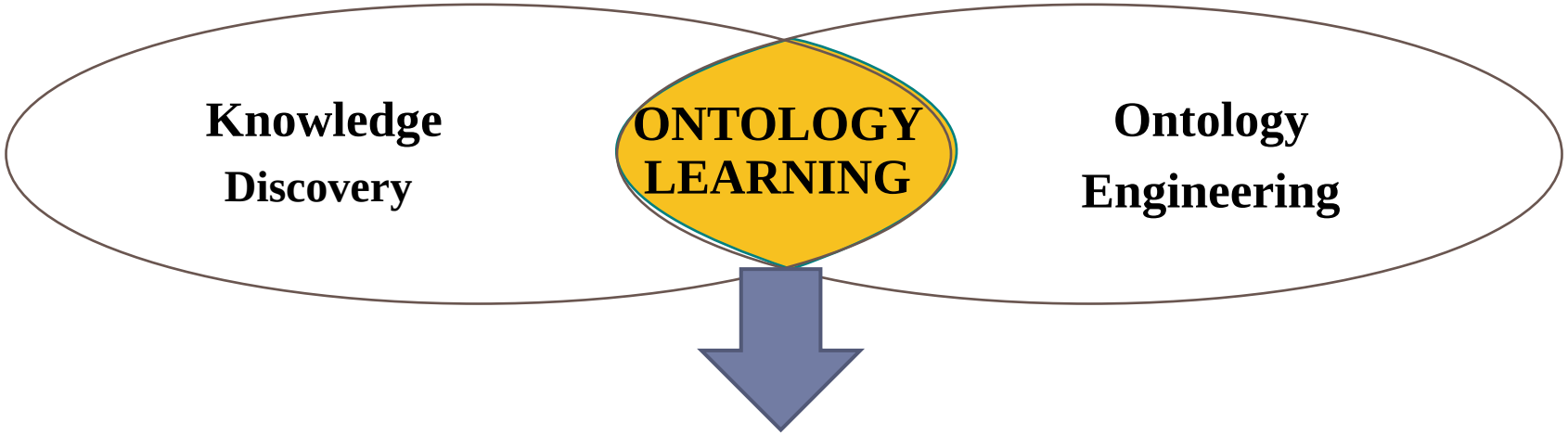


Manual Ontology engineering

- Manual ontology engineering:
  - Ontology building « from scratch » [Fernandez et al., 99]
  - OTK Methodology: Knowledge Meta Process [Y. Sure and R. Studer, 2002]
  - Methontology [Fernández-Lopez et al., 1997]
- But….
  - Manual engineering of ontologies is a very time consuming task!

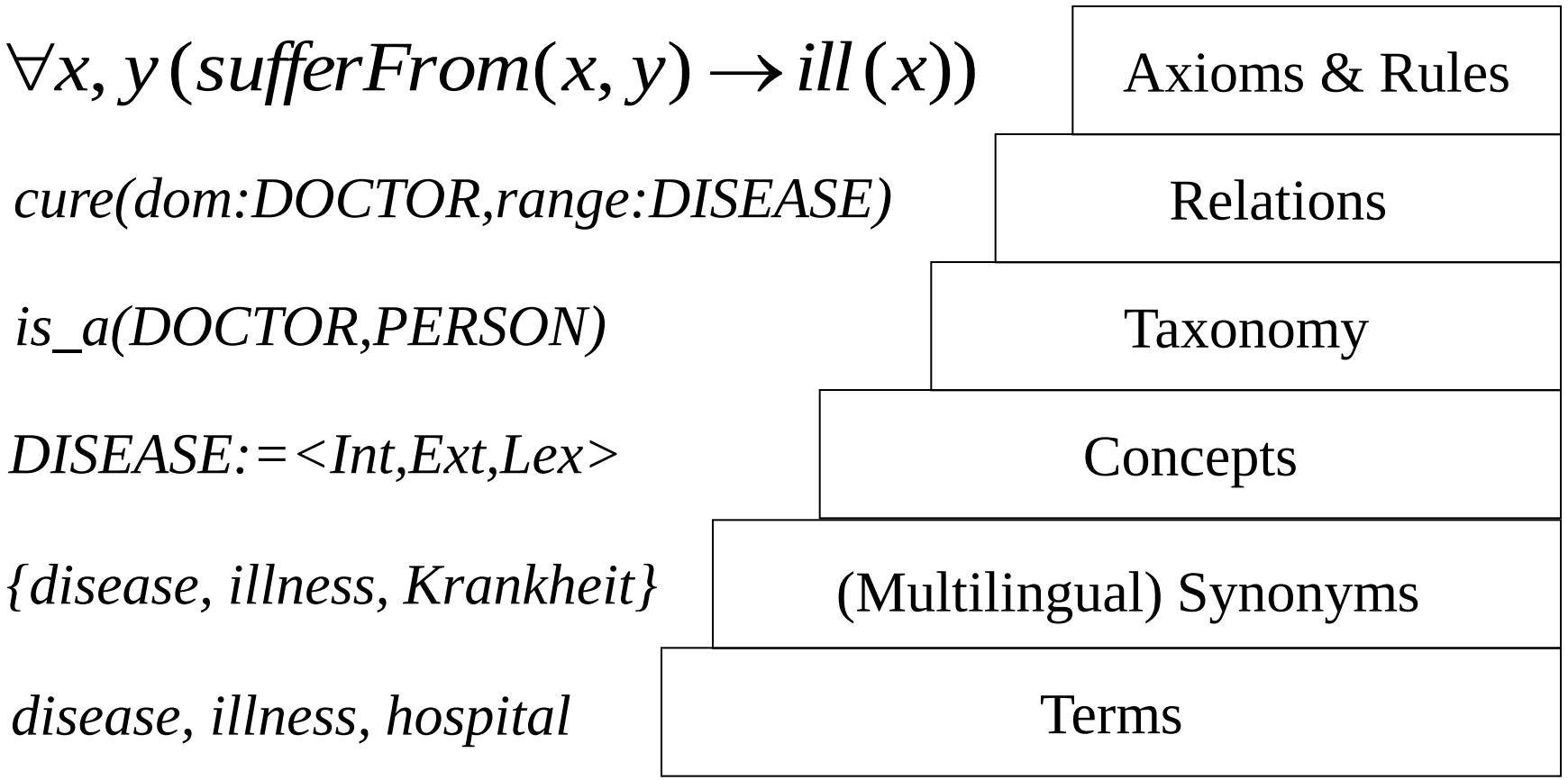
=> **Need of automatic way to reduce the burden of engineering !**

Domain Ontology Learning



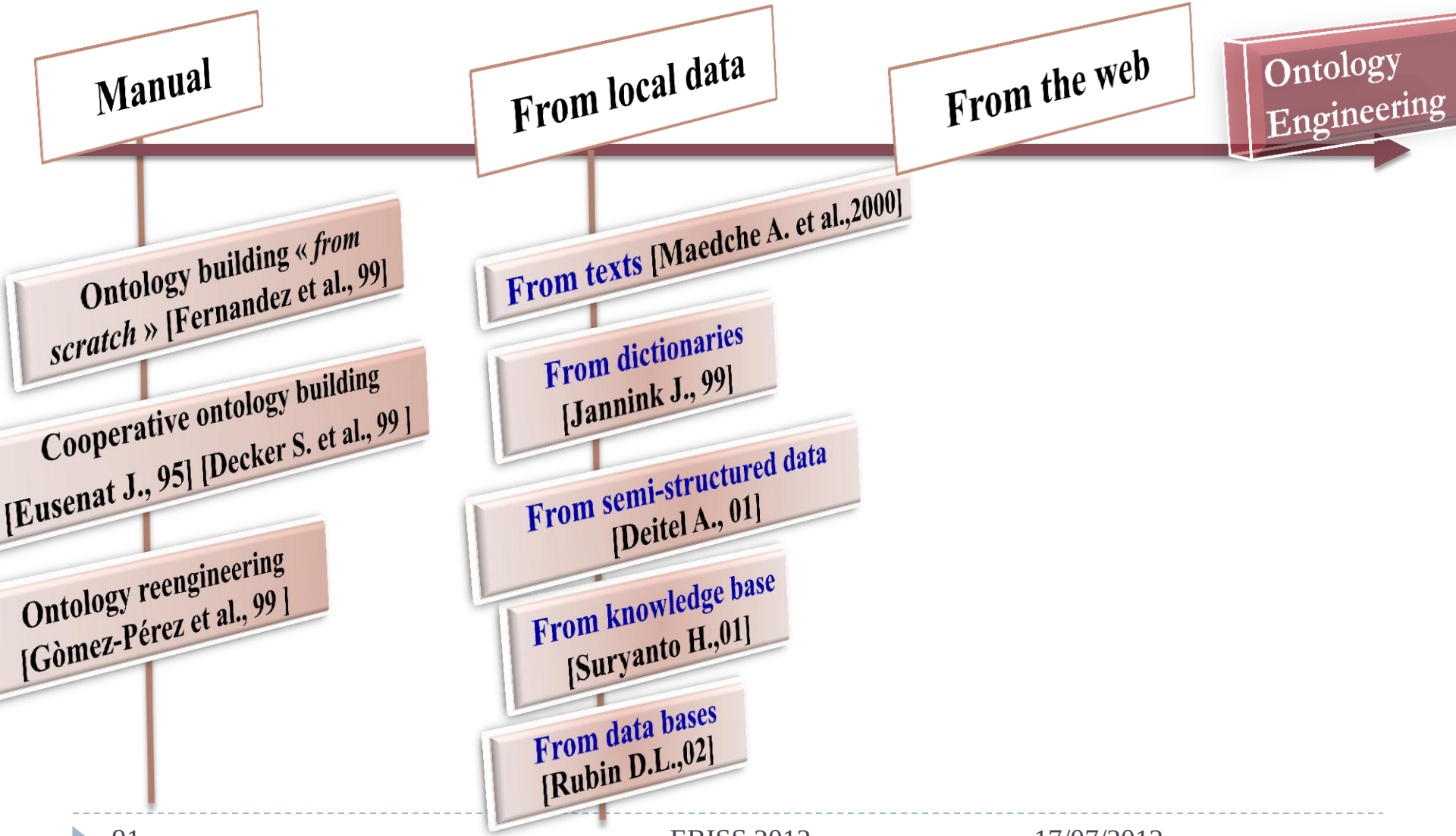
Ontology learning *is defined as an approach of ontology building from knowledge sources using a set of* ***machine learning techniques*** *and knowledge* ***acquisition methods***.

DomainOntology Learning
Main steps



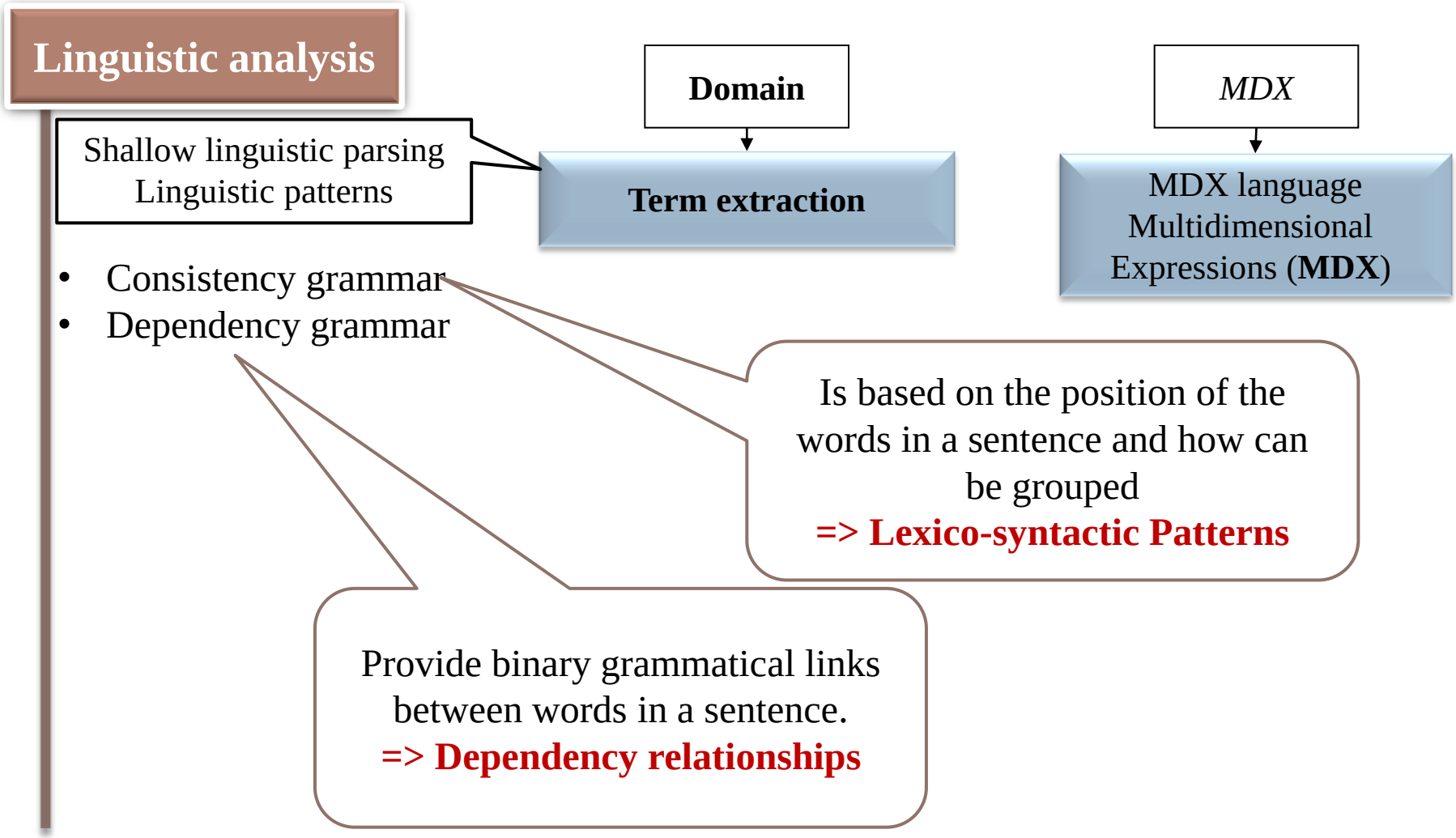
Introduced in: Philipp Cimiano, PhD Thesis University of Karlsruhe, forthcoming

Domain Ontology Learning



Domain Ontology learning from Texts

Domain Ontology learning from Texts



Domain Ontology learning from Texts
Term Extraction

Text: *Multidimensional Expressions (**MDX**) **is a** query language for OLAP databases*

Linguistic analysis

- Consistency grammar

```
(ROOT
  (S
    (NP
      (NP (NNP Multidimensional) (NNP Expressions))
      (PRN (-LRB- -LRB-)
        (NP (NNP MDX))
        (-RRB- -RRB-)))
    (VP (VBZ is)
      (NP
        (NP (DT a) (JJ query) (NN language))
        (PP (IN for)
          (NP (NNP OLAP) (NNS databases)))))))
```

- Dependency grammar

```
nn(Expressions-2, Multidimensional-1)
nsubj(language-9, Expressions-2)
abbrev(Expressions-2, MDX-4)
cop(language-9, is-6)
det(language-9, a-7)
amod(language-9, query-8)
root(ROOT-0, language-9)
prep(language-9, for-10)
nn(databases-12, OLAP-11)
pobj(for-10, databases-12)
```

Domain Ontology learning from Texts

Linguistic analysis

Shallow linguistic parsing & Linguistic patterns

- Consistency grammar
- Dependency grammar

Statistical analysis

Co-occurrence analysis

- Selection measures (IR)

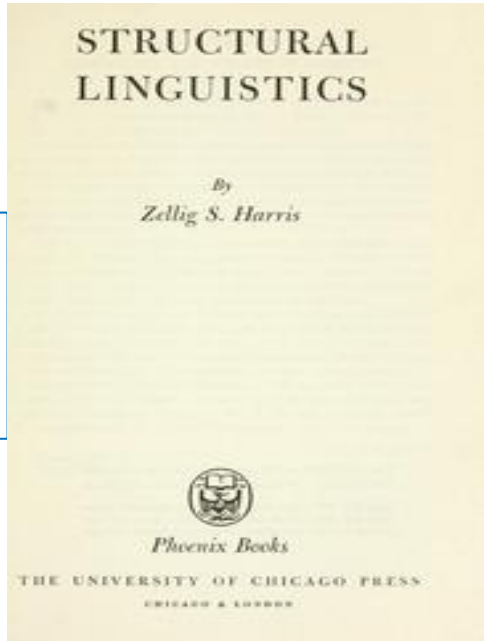
Domain

Term extraction

MDX

MDX language Multidimensional Expressions (**MDX**)

Methods that determine a score representing the relationship between two terms and retain those with scores greater than a given threshold



Domain Ontology learning from Texts
Term Extraction

Statistical analysis

TF-IDF [Robertson 1976]

Term_frequency(t,d) * Inverse_term_frequency(t, D)

Entropy [Brini 2005]

$$-\sum_{t \in Di}^{Dn} P(t)\ log\ (P(t))$$

Relevance to the domain

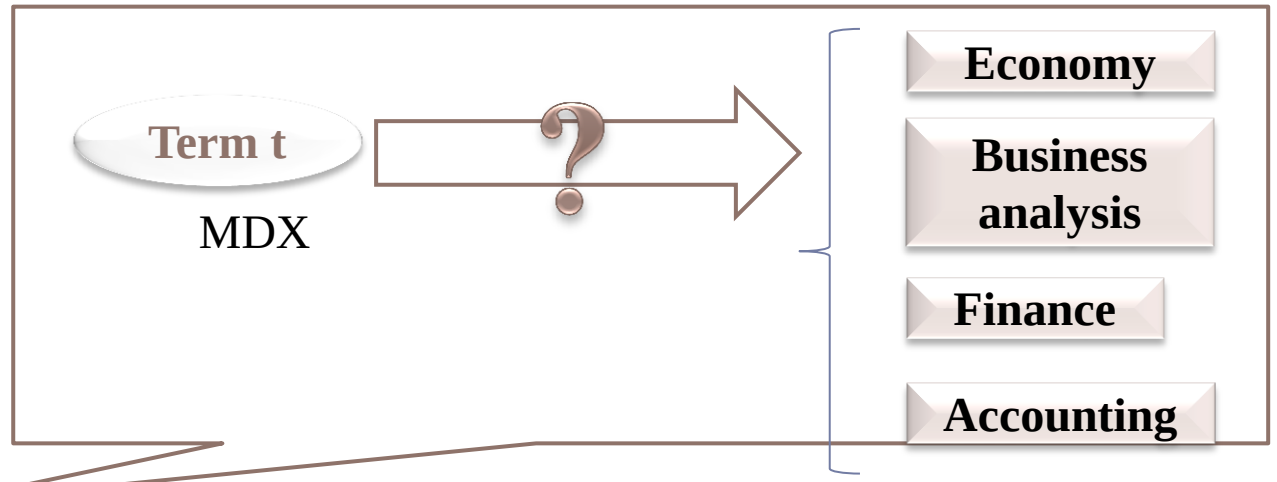
Consensus to the domain

Pointwise Mutual Information (PMI)

Domain Ontology learning from Texts
Term Extraction

Statistical analysis

- **TF-IDF [Robertson 1976]**
- **Entropy [Brini 2005]**
- **Relevance to the domain [Verladi, 2001**
- **Consensus to the domain**
- **Pointwise Mutual Information (PMI)**



$$RD(t, domain_i) = \frac{P(t, domain_i)}{\sum_{i=1..n} p(t, domain_i)}$$

$$P(t, domain_i) = \frac{freq(t\ in\ domain_i)}{\sum_{i=1..n} p(t, domain_i)}$$

Domain Ontology learning from Texts
Term Extraction

Statistical analysis

- **TF-IDF [Robertson 1976]**
- **Entropy [Brini 2005]**
- **Relevance to the domain [Verladi, 2001]**
- **Consensus to the domain**
- **Pointwise Mutual Information (PMI)**

Analysis of the distribution of the term t on a set of documents dj belonging to the domain Di.

$$CD(t, D_i)=\sum P(t, dj)\ \log 2(\frac{1}{P(t,dj)})$$

$$P(t, D_i)=\frac{freq(t\ in\ \mathrm{dj})}{\sum_{i=1..n} p(t\ in\ \mathrm{d}j)}$$

Domain Ontology learning from Texts
Term Extraction

Statistical analysis

- **TF-IDF [Robertson 1976]**
- **Entropy [Brini 2005]**
- **Relevance to the domain [Verladi, 2001]**
- **Consensus to the domain**
- **Pointwise Mutual Information (PMI)**

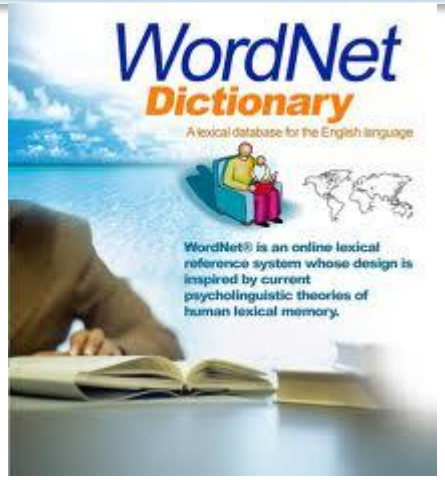
Probability is estimated to the co-occurrence of the term t with the domain concepts "concept"

$$PMI(t, concept)=\log\left(\frac{p(t/concept)}{p(concept)}\right)$$

Domain Ontology learning from Texts
Concept Discovery

Hybrid methods

Lexical resources:
Wordnet



Domain

MDX

WordNet 2.1 Browser

File History Options Help

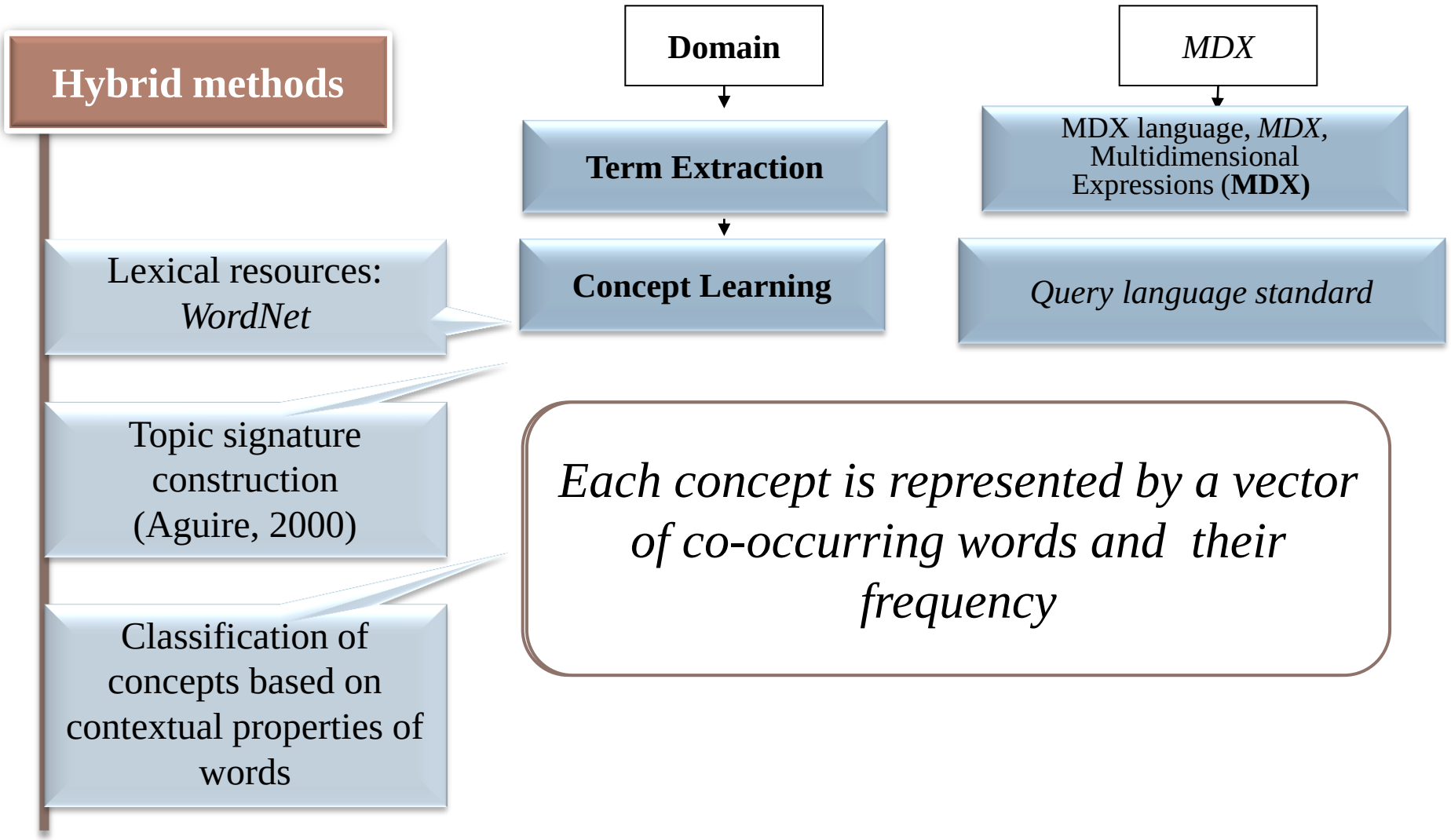
Search Word: language

Searches for language: Noun Senses:

The noun language has 6 senses (first 6 from tagged texts)

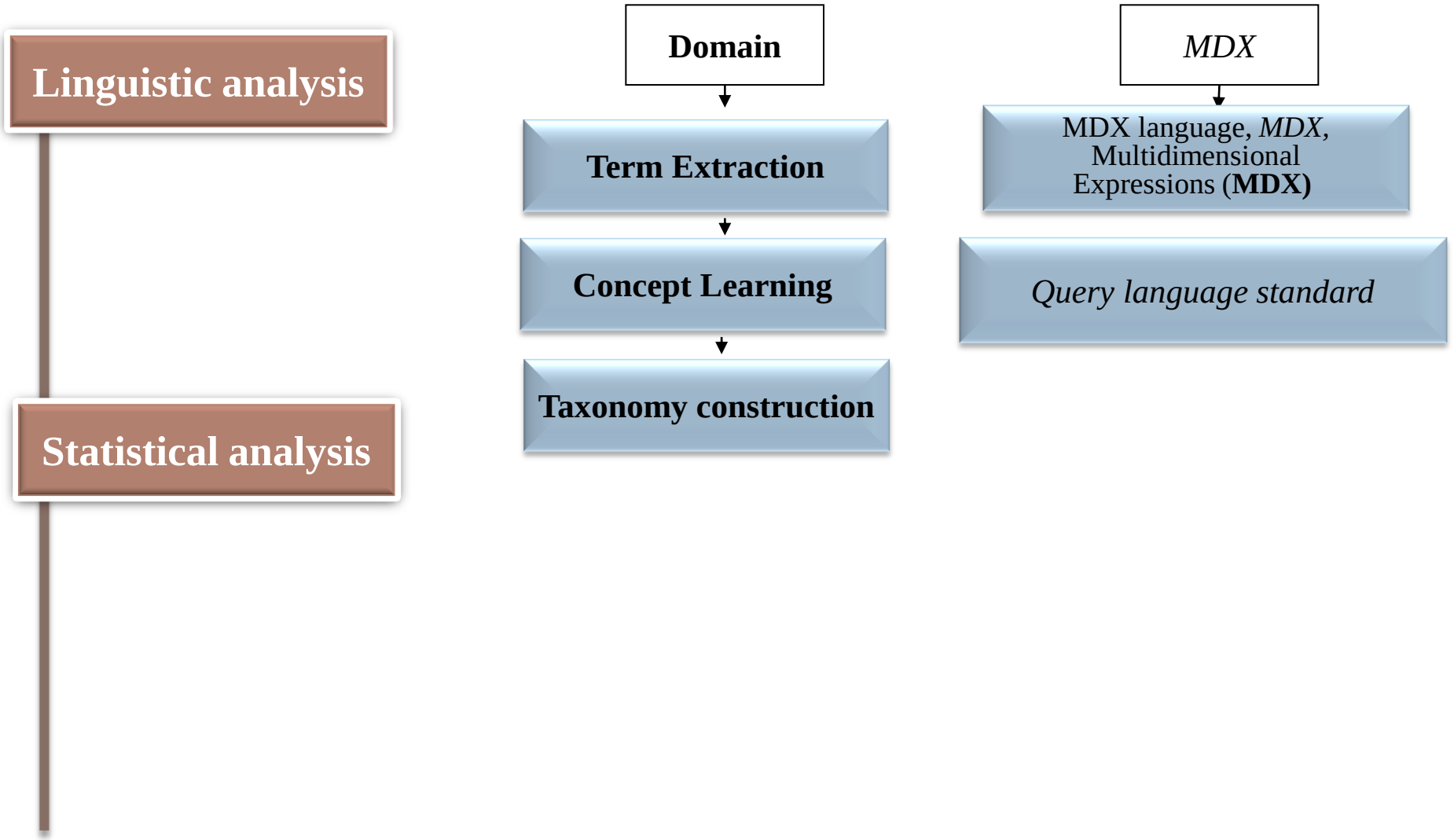
1. (486) **language**, linguistic communication -- (a systematic means of communicating by the use of sounds or conventional symbols; "he taught foreign languages"; "the language introduced is standard throughout the text"; "the speed with which a program can be executed depends on the language in which it is written")
2. (155) speech, speech communication, spoken communication, spoken language, **language**, voice communication, oral communication -- ((language) communication by word of mouth; "his speech was garbled"; "he uttered harsh language"; "he recorded the spoken language of the streets")
3. (28) terminology, nomenclature, **language** -- (a system of words used to name things in a particular discipline; "legal terminology"; "biological nomenclature"; "the language of sociology")
4. (25) linguistic process, **language** -- (the cognitive processes involved in producing and understanding linguistic communication; "he didn't have the language to express his feelings")
5. (13) **language**, speech -- (the mental faculty or power of vocal communication; "language sets homo sapiens apart from all other animals")
6. (5) lyric, words, **language** -- (the text of a popular song or musical-comedy number; "his compositions always started with the lyrics"; "he wrote both words and music"; "the song uses colloquial language")

Domain Ontology learning from Texts
Concept Discovery



Domain Ontology learning from Texts

Taxonomy Construction



Domain Ontology learning from Texts
Taxonomy Construction

Linguistic analysis

Lexico-syntactic patterns (Hearst, 1998)

NP such as NP, NP, ... and NP

*a multidimensional database **query language** **such as** **MDX**.*

Such NP as NP, NP, ... or NP

Such** **query language** **as** **MDX…….

NP, NP, ... and other NP

*screen real estate to financial charts**,** indices **and other** news graphics.*

NP, especially NP, NP,... and NP

*Accounting, **especially** financial accounting gives mainly past information in that the events are recorded.*

Domain Ontology learning from Texts
Taxonomy Construction

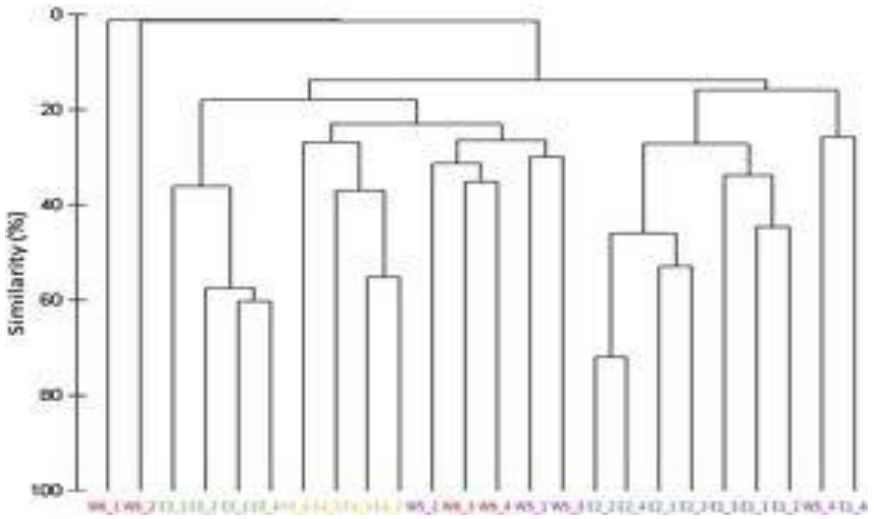
Linguistic analysis

Lexico-syntactic patterns (Hearst, 1998)

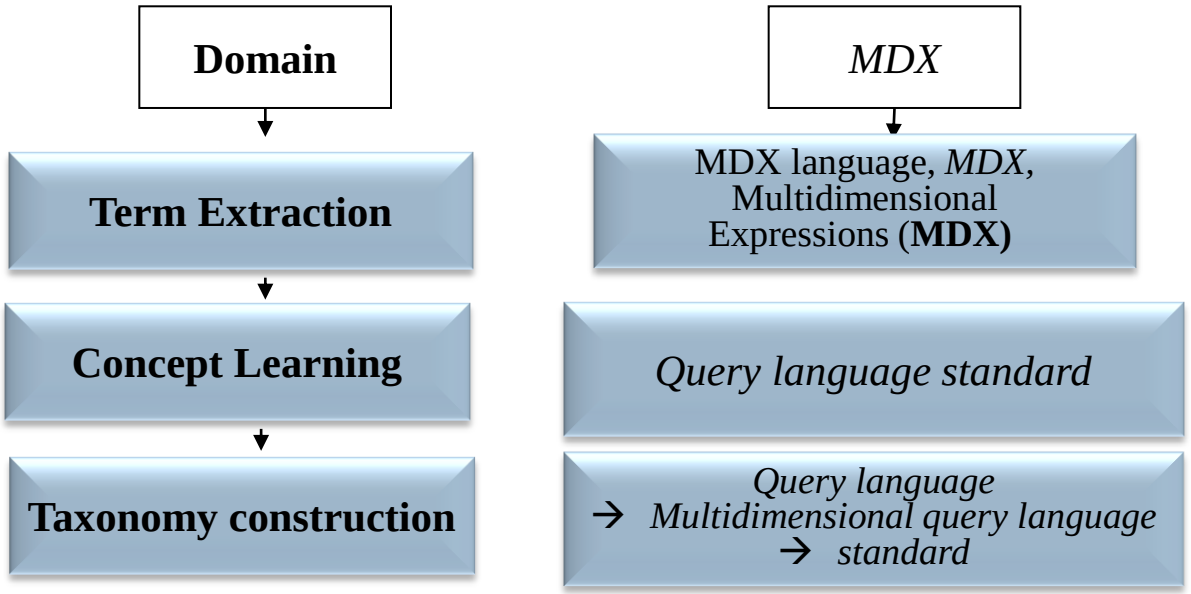
Statistical analysis

Hierarchical grouping of concepts

Grouping based on probability-based measures

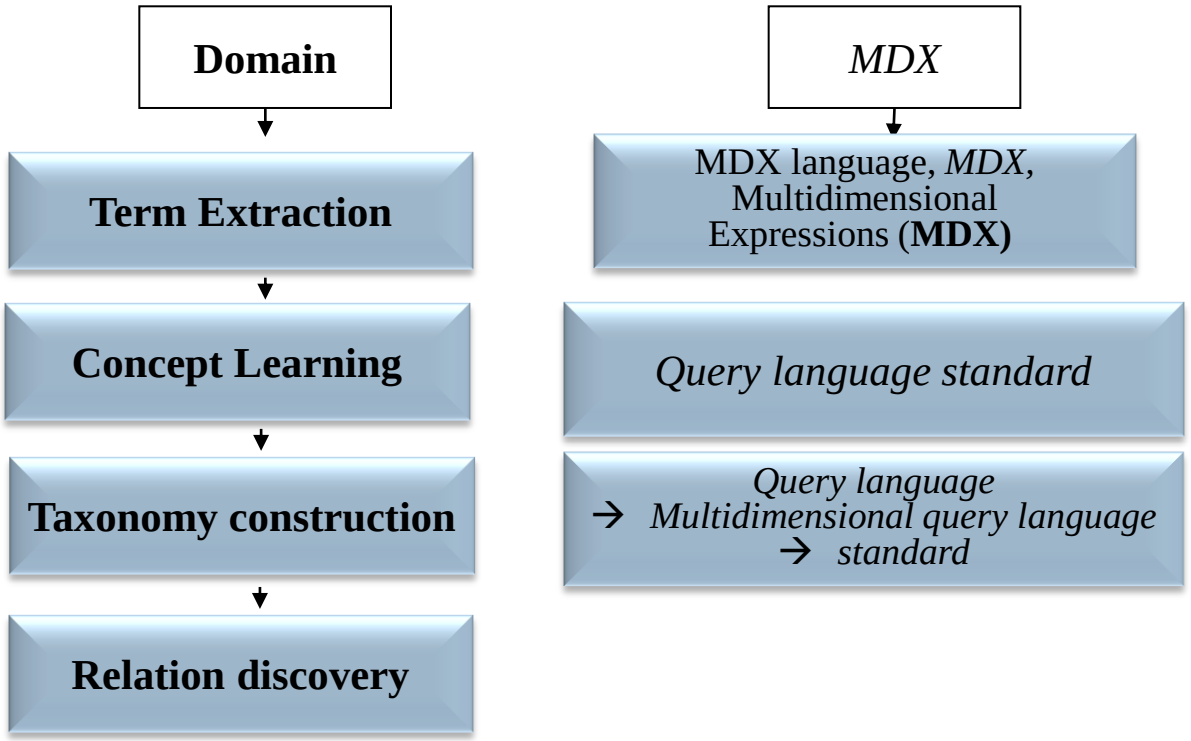


Domain Ontology learning from Texts



Domain Ontology learning from Texts

Relation Discovery



Domain Ontology learning from Texts
Relation Discovery

Linguistic analysis

Syntactic frames *[Faure and Nedellec, 1998]*

<To be used > (<subject: MDX>) (<for: OLAP database analysis>) <To be used > (<subject: MDX>) (< for : OLAP database querying >);

Statistical analysis

Conceptual clustering

Multidimensional query language — Is used for → *OLAP Database management*

Domain Ontology learning from Texts
Relation Discovery

Linguistic analysis

Syntactic frames
[Faure and Nedellec, 1998]

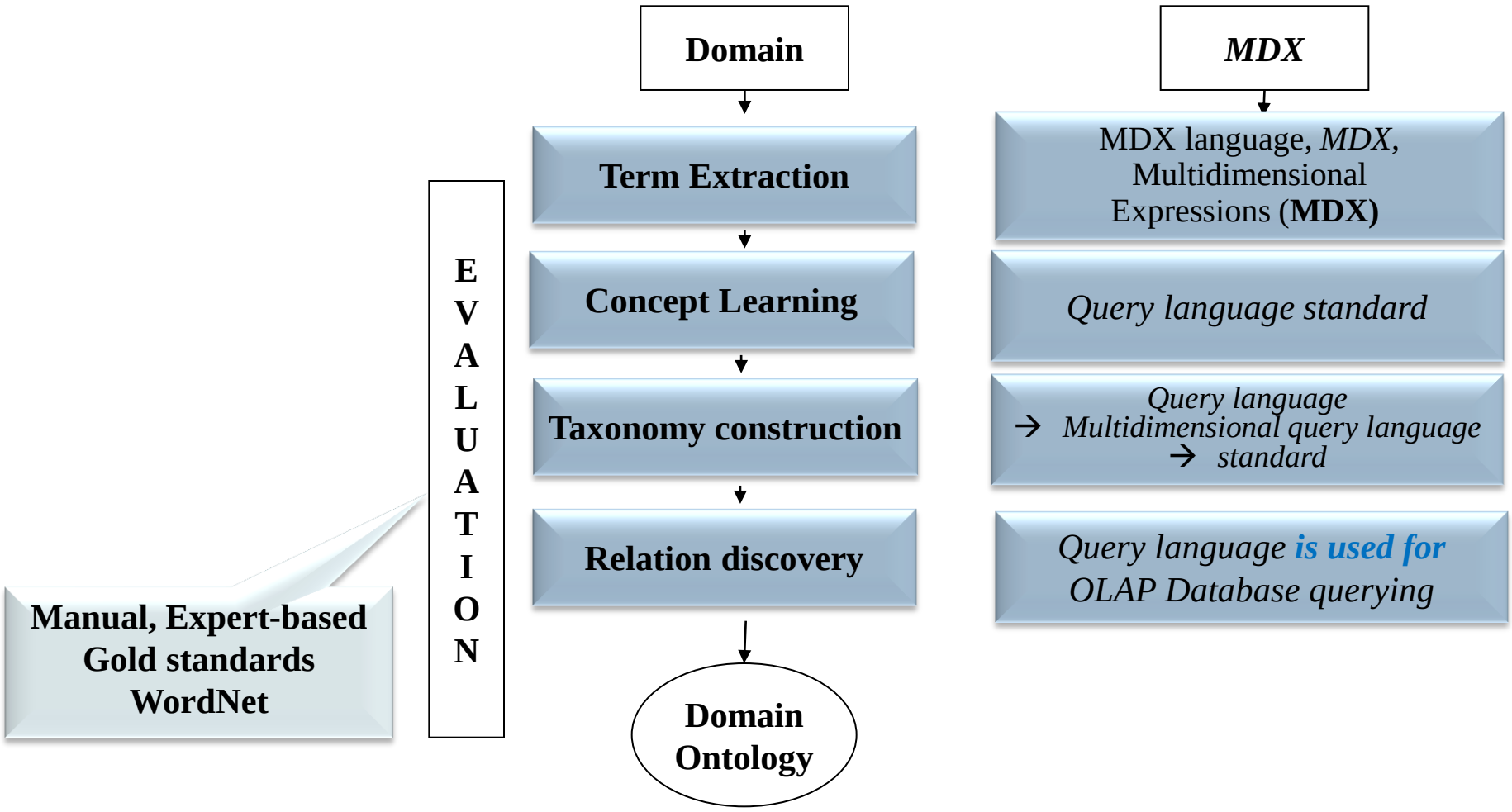
Statistical analysis

Association rules
Co-occurrence analysis
Verb arguments analysis

	w1	w2	w3		Wn-1	wn
w1						
w2						
w3						
Wn-1						
wn						

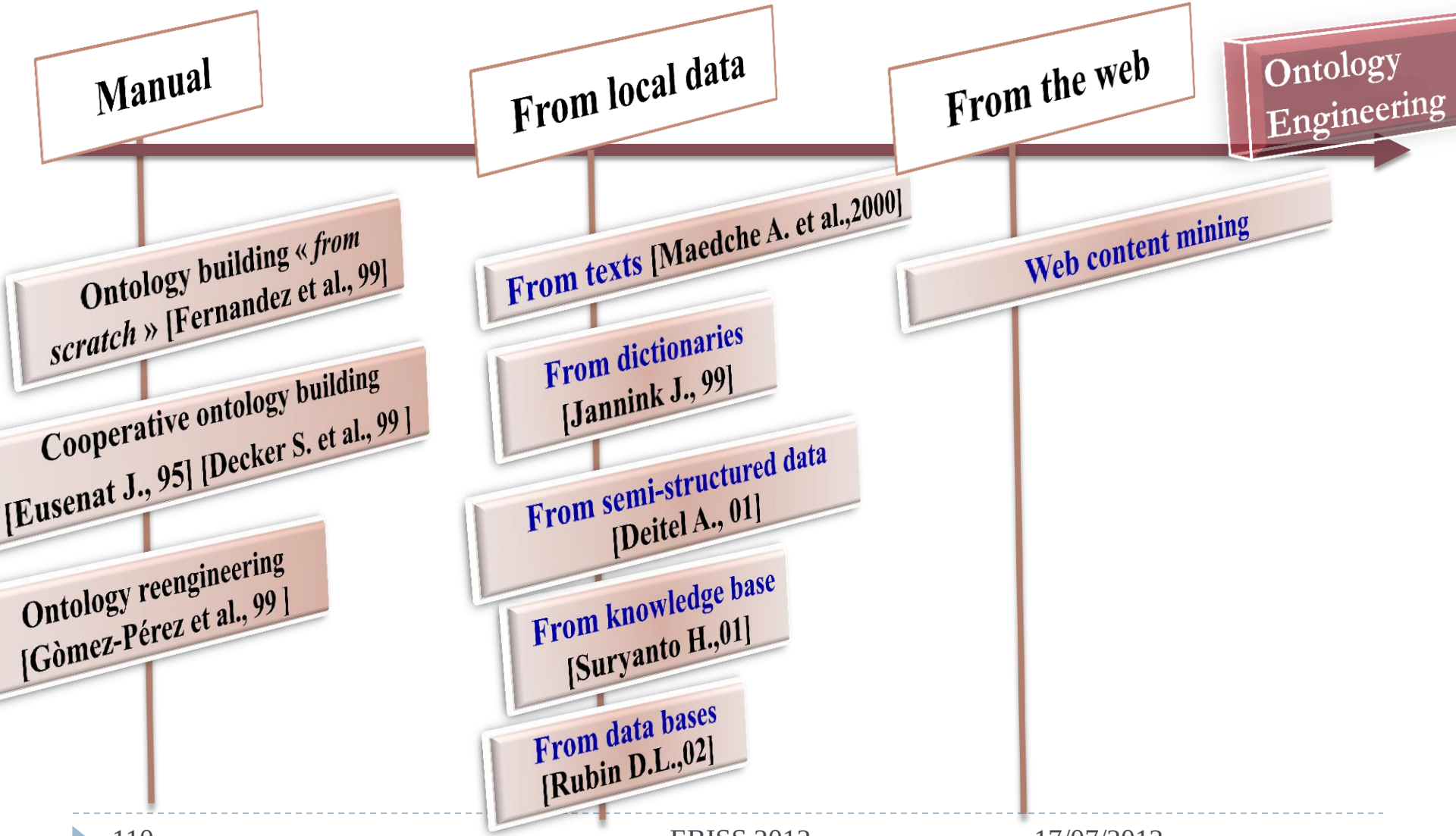
Word space
DOODLE II[**Sugiura et al., 2003**]

Domain Ontology learning from Texts



Domain Ontology Learning

The Web as a learning source….

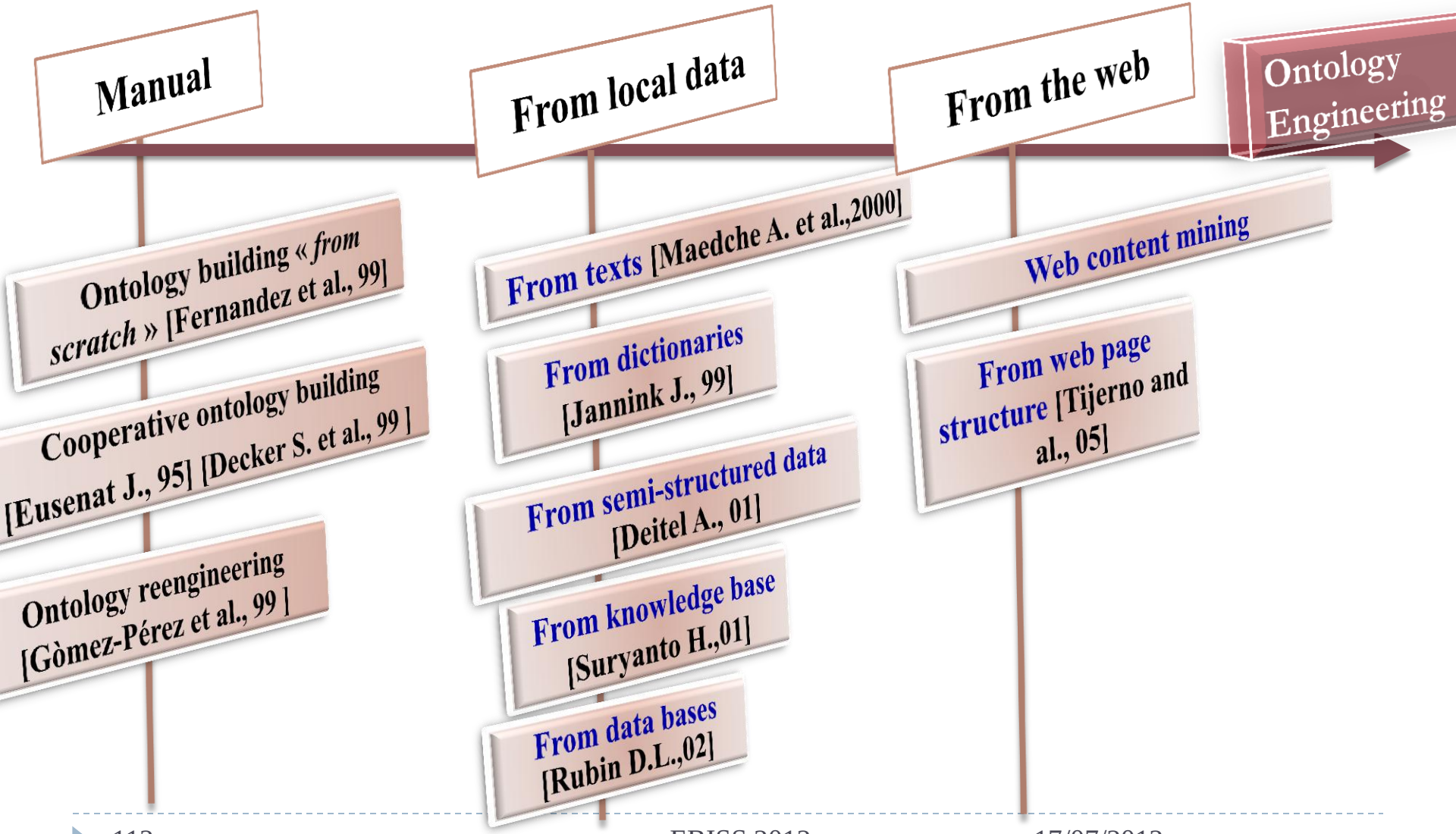


Domain Ontology Learning
From web pages (texts)

- Collection of web pages related to a specific domain
- Application of the same techniques as textual approaches
- Web: Largest electronic textual corpora

- Shortcomings
  - *NL resources, visual-oriented representation*
  - *Noise (commercial bias)*
  - *Lack of semantic structure*
  - *Unreliable source*
- Advantages
  - *Size and heterogeneity*
  - ***Public massive IR tools***

Domain Ontology Learning



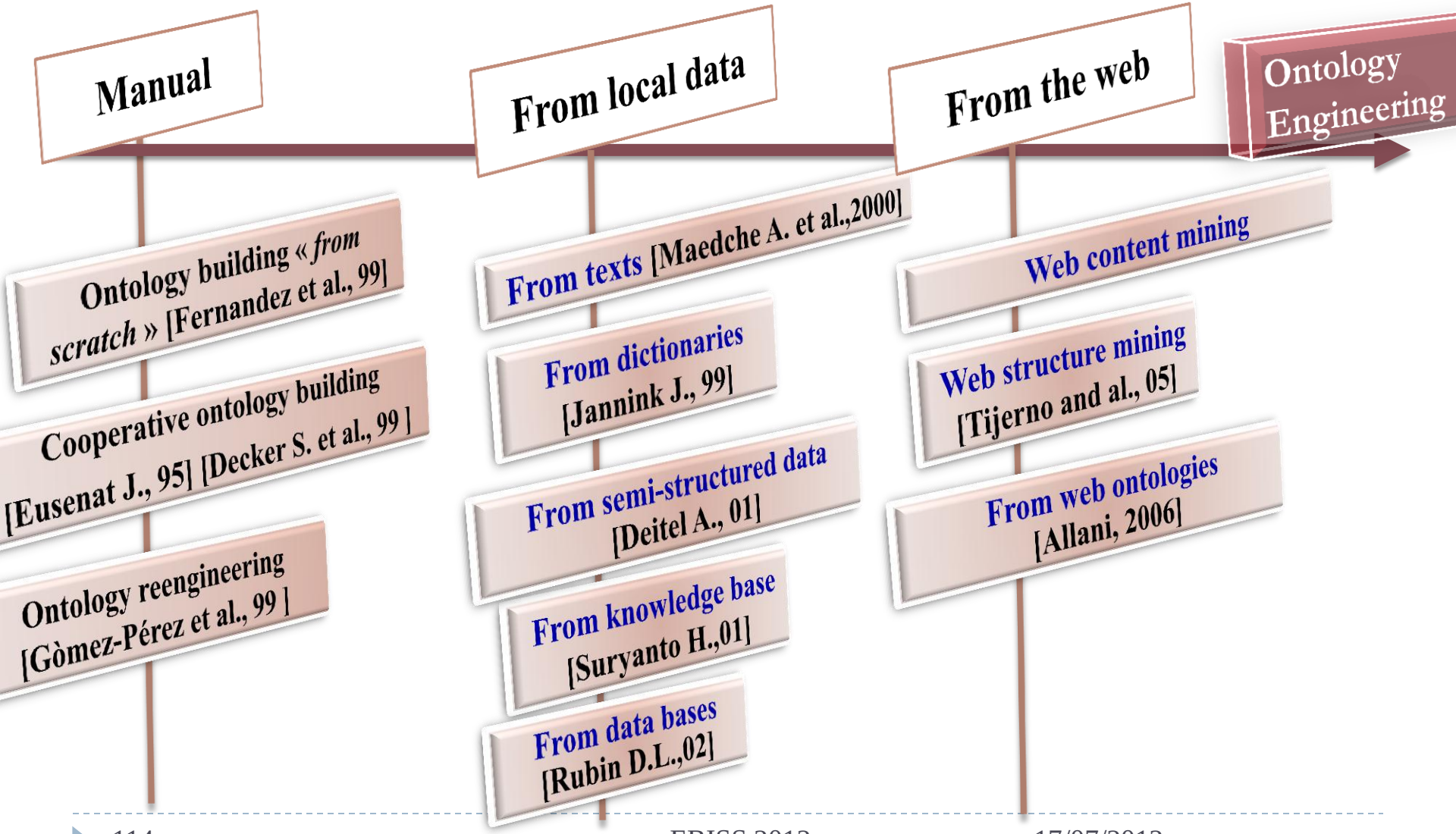
Ontology Learning from web
From web page structure

- Nouns phrases in the headings, tables or lists of a document can be used to deduce relations between concepts and instances

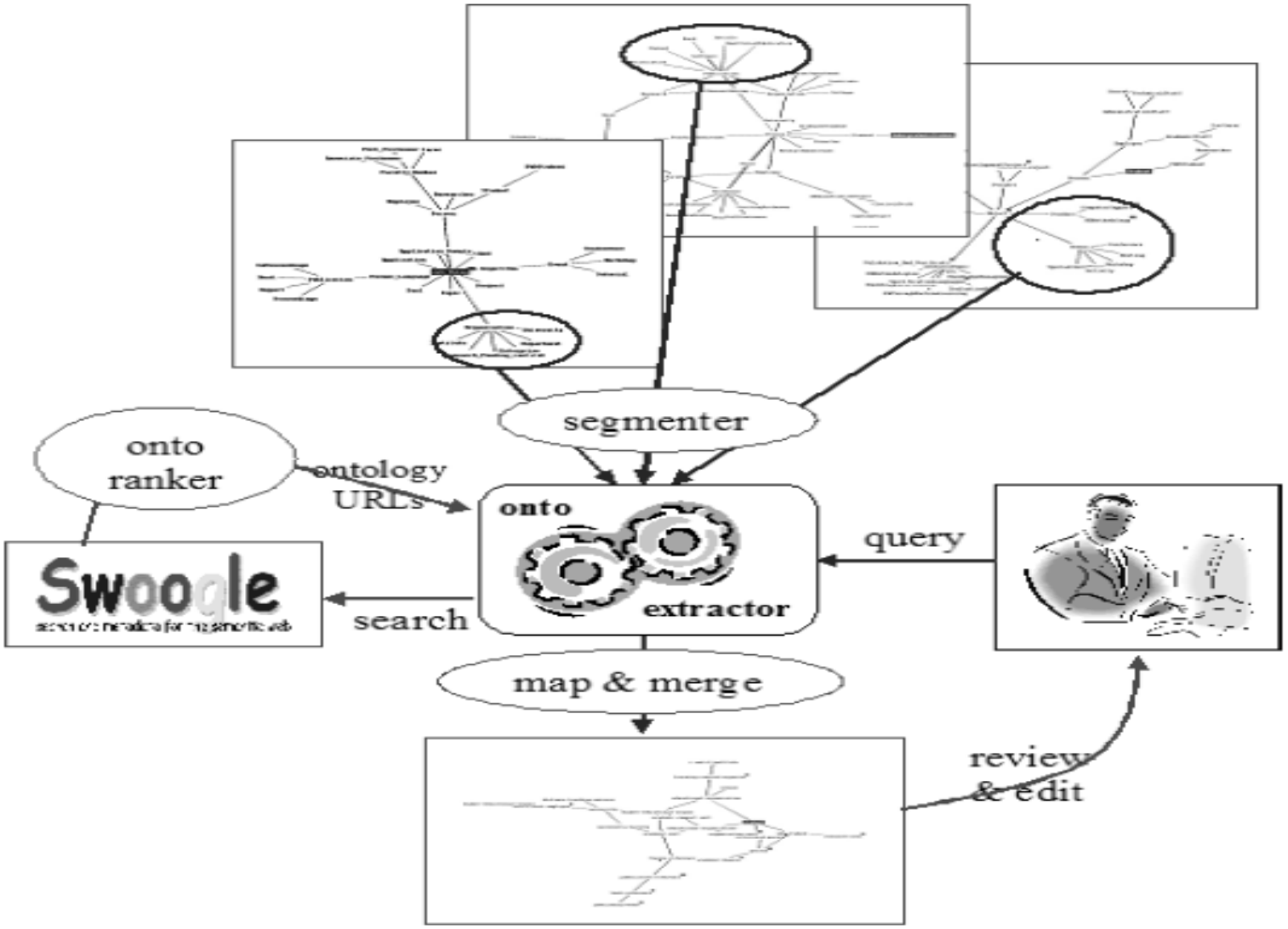
Business District	Office Space (m²)	Business Concentration
The City	7,740,000	finance, broking, insurance, legal
Westminster	5,780,000	head offices, real estate, private banking, hedge funds, government
Camden & Islington	2,294,000	creative industries, finance, design, art, fashion, architecture
Canary Wharf	2,120,000	banking, media, legal
Lambeth & Southwark	1,780,000	accountancy, consultancy, local government

☹ Supervised approaches
☹ Difficulty of the interpretation of html structure

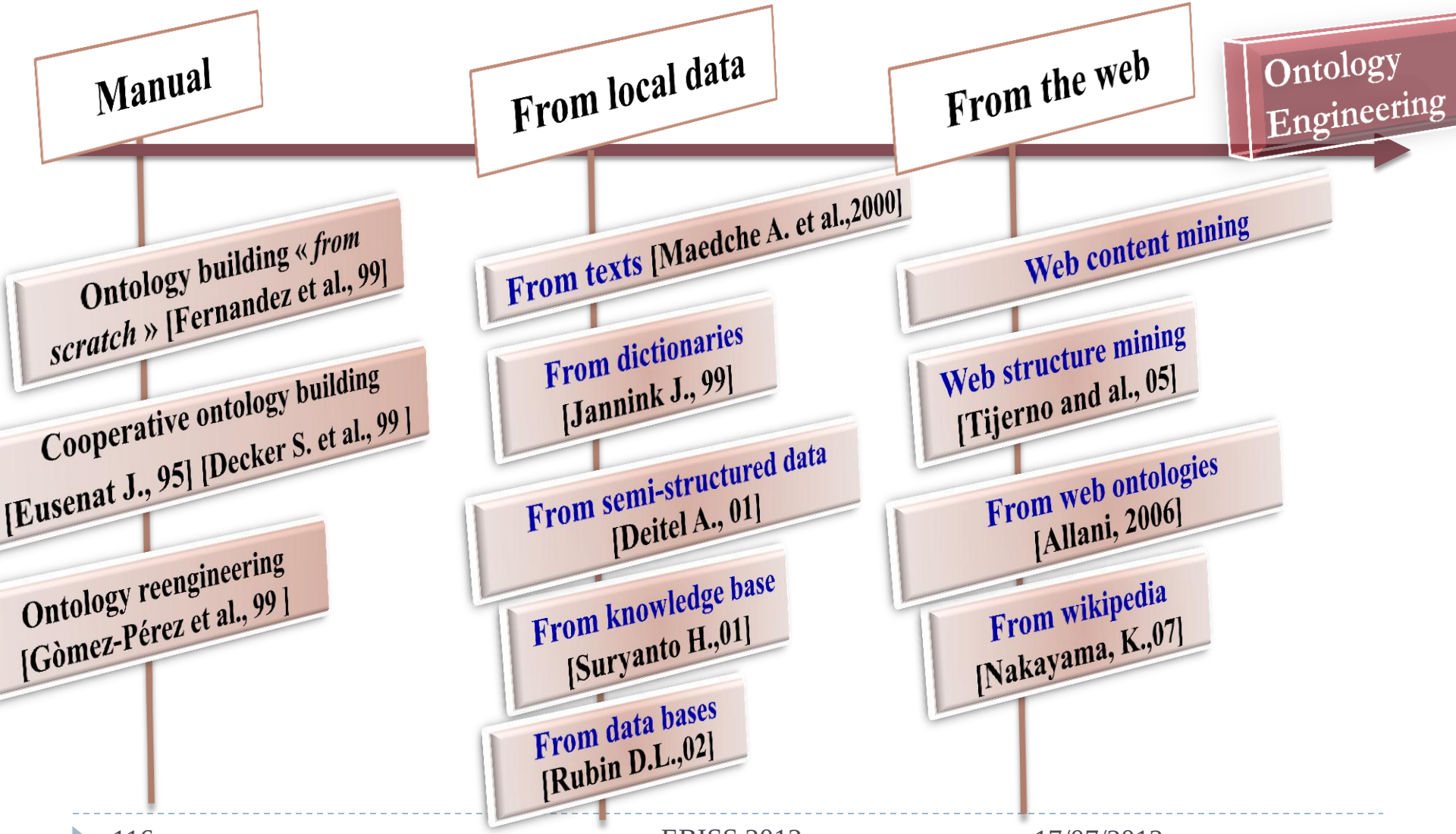
Domain Ontology Learning



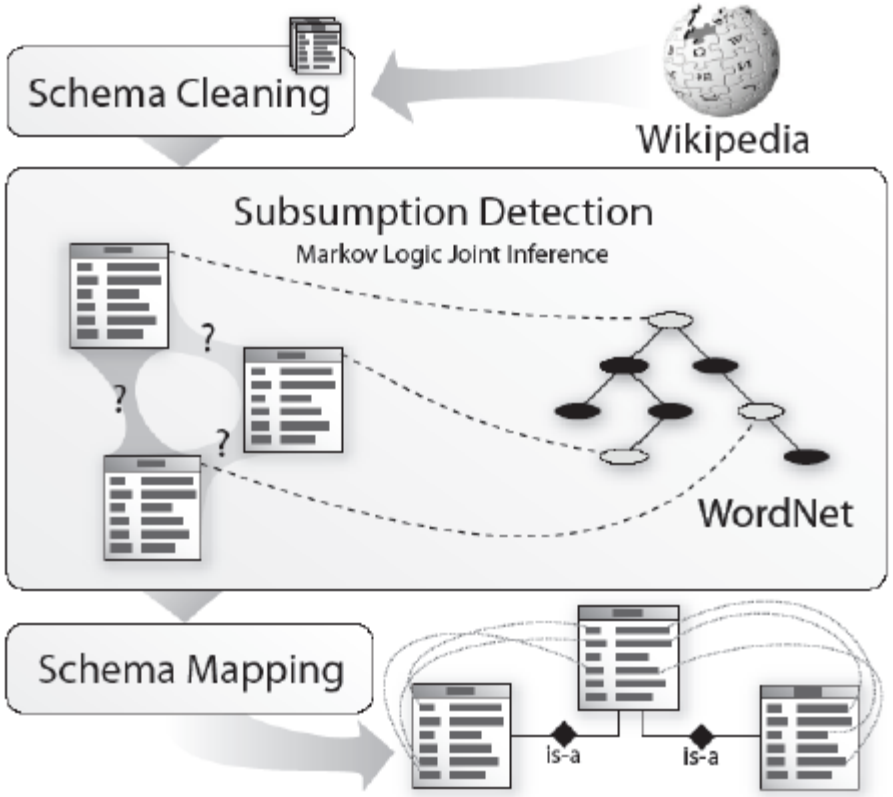
Ontology Learning from web
Aggregation of on-line ontologies

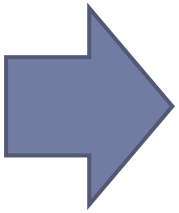


Domain Ontology Learning



Ontology Learning from web
Wikipedia mining

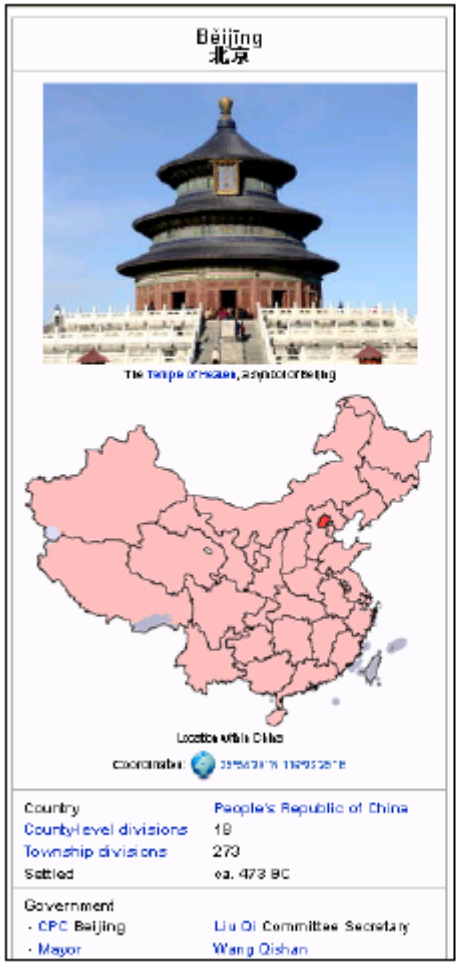




YAGO Ontology

Wu and Weld, 2008

Ontology Learning from web
Wikipedia mining



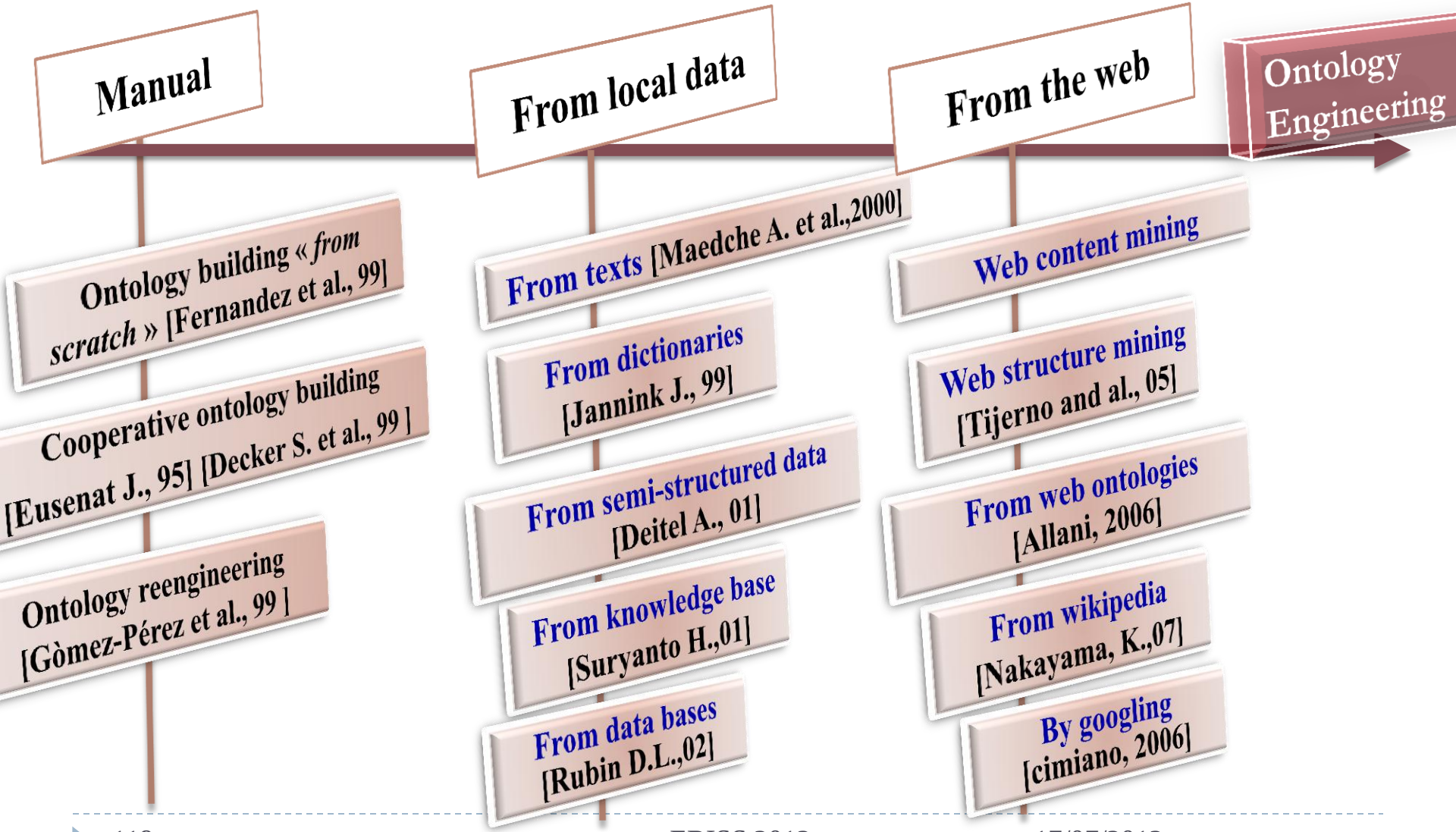

```
{{Infobox Settlement
|official_name     = Běijīng
|other_name        =
|native_name       = 北京
|settlement_type   = [[Municipality of
                     China|Municipality
|image_skyline     = SA Temple of Heaven.jpg
|image_caption     = The [[Temple of
                     Heaven]], a symbol of Beijing
|citylogo_size     =
|image_map         = China-Beijing.png
|mapsize           = 275px
|map_caption       = Location within China
|subdivision_type  = Country
|subdivision_name  = [[People's Republic of China]]
|subdivision_type1 = [[Political divisions of
                     China#County level|County-
                     level divisions]]
|subdivision_name  = 18
|subdivision_type2 = [[Political divisions of
                     China#Township
                     level|Township divisions]]
|subdivision_name2 = 273
|leader_title      =[[Communist Party of
                     China|CPC]] Beijing
|leader_name       =[[Liu Qi (Communist)|Liu Qi]]
                     Committee Secretary
|leader_title1     = [[Mayor]]
|leader_name1      =[[Wang Qishan]]
|established_title = Settled
|established_date  = ca. 473 BC
...
}}
```

☹ terminological ontology

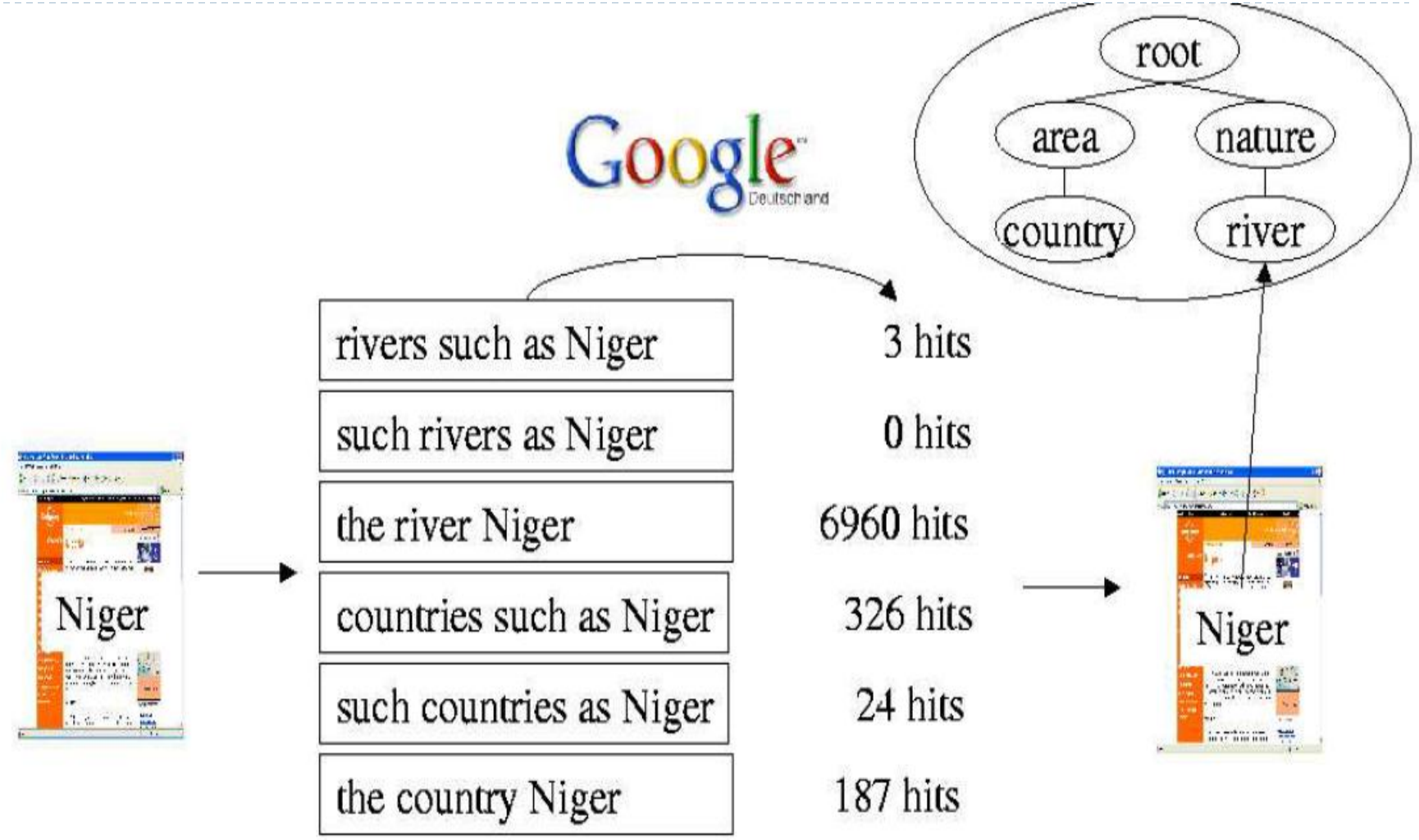
☹ not scalable

Extraction of schema from infoboxes (Wu and Weld, 2008)

Domain Ontology Learning



Ontology Learning from web
Ontology learning by Googling



Unsupervised Web-based semantic relatedness for ontology learning

- The correlation function C_k (a, b) [Turney, 2001] = $\frac{p(ab)^k}{p(a)p(b)}$

- Symmetric Conditional Probability (SCP) [ferreira da Silva and lopes, 1999]

SCP (a,b) = C_2 (a, b)= $\frac{p(a,b)^2}{p(a) * p(b)}$

- The point wise mutual information (PMI) [church and al., 1991]

PMI(a,b) = $\log_2 \frac{p(ab)}{p(a)\,p(b)}$

- The correlation function C_k (a, b) [Turney, 2001] = $\frac{p(ab)^k}{p(a) * p(b)}$

If the words a and b are independent → p(a and b) = p(a) * p(b)

Else p(a and b) > p(a) * p(b)

For concept selection:

$$c(problem, choice) = \frac{p(problem, choice)}{p(choice)} \quad (2)$$

- Web = scalable Corpora
- P (problem, choice) = $\frac{hit(problem\ AND\ choice)}{Total_pages_Web}$

$$c(problem, choice) = \frac{hits(problem\ AND\ choice)}{hits(choice)} \quad (3)$$

Ontology Learning from web
Ontology learning by Googling

* "such as financial accounting" *

Environ 3 450 000 résultats (0,26 secondes)

[PDF] A4 6pp
www2.lse.ac.uk/.../pdf/Diploma%20brochure.pdf - Traduire cette page
Format de fichier: PDF/Adobe Acrobat - Afficher
topics such as financial accounting, managerial accounting, auditing, asset pricing, corporate finance, and valuation. Students can also acquire knowledge of, or ...

[PDF] THE RELATIONSHIP BETWEEN BUSINESS FINANCE AND ...
www.leidykla.eu/fileadmin/.../Jana_Kajanova.pdf - Traduire cette page
Format de fichier: PDF/Adobe Acrobat - Afficher
All indicated subsystems of the **accounting system, such as financial accounting, managerial** accounting, intraorganizational accounting and tax accounting do ...

ESS-MSS Introduction Class-1-SU - YouTube

www.youtube.com/watch?v=2Q8TYEhfcqo
17 juin 2012 - 75 min - Ajouté par varunsas09
... Malaysia, China, Thailand and Australia for **SAP functional skills such as financial accounting and** ...

How to become CPA - YouTube

www.youtube.com/watch?v=Kvca8dTuBug
20 déc. 2011 - 6 min - Ajouté par IPassTheCPAExam
... that is, auditing experience while others are fine with non-public **accounting work such as financial** ...

Ontology Learning from web
Ontology learning by Googling

- Non taxonomic relation learning
  - Extraction verb phrases from snippets

score (verb_phrase, concept) = $$\frac{\text{hit}(\text{“verb_phrase concept”})}{\text{hit}(\text{verb_phrase})}$$

  - Retrieval and selection of related concepts

score (candidate, concept)= $$\frac{hit(\text{"}concept\text{"}\ AND\ \text{"}candidate\text{"})}{hit(concept)}$$

Unsupervised Web-based semantic relatedness for ontology learning : problems

- $$c(concept, candidate) = \frac{hits(concept\ AND\ candidate)}{hits(candidate)}$$

- Language ambiguity
- Polysemy
- hits (word) = Somme (hits ($sens_1$), hits ($sens_2$), …., hits($sens_n$))

→ Need to contextualize these measures (Introducing context of words)

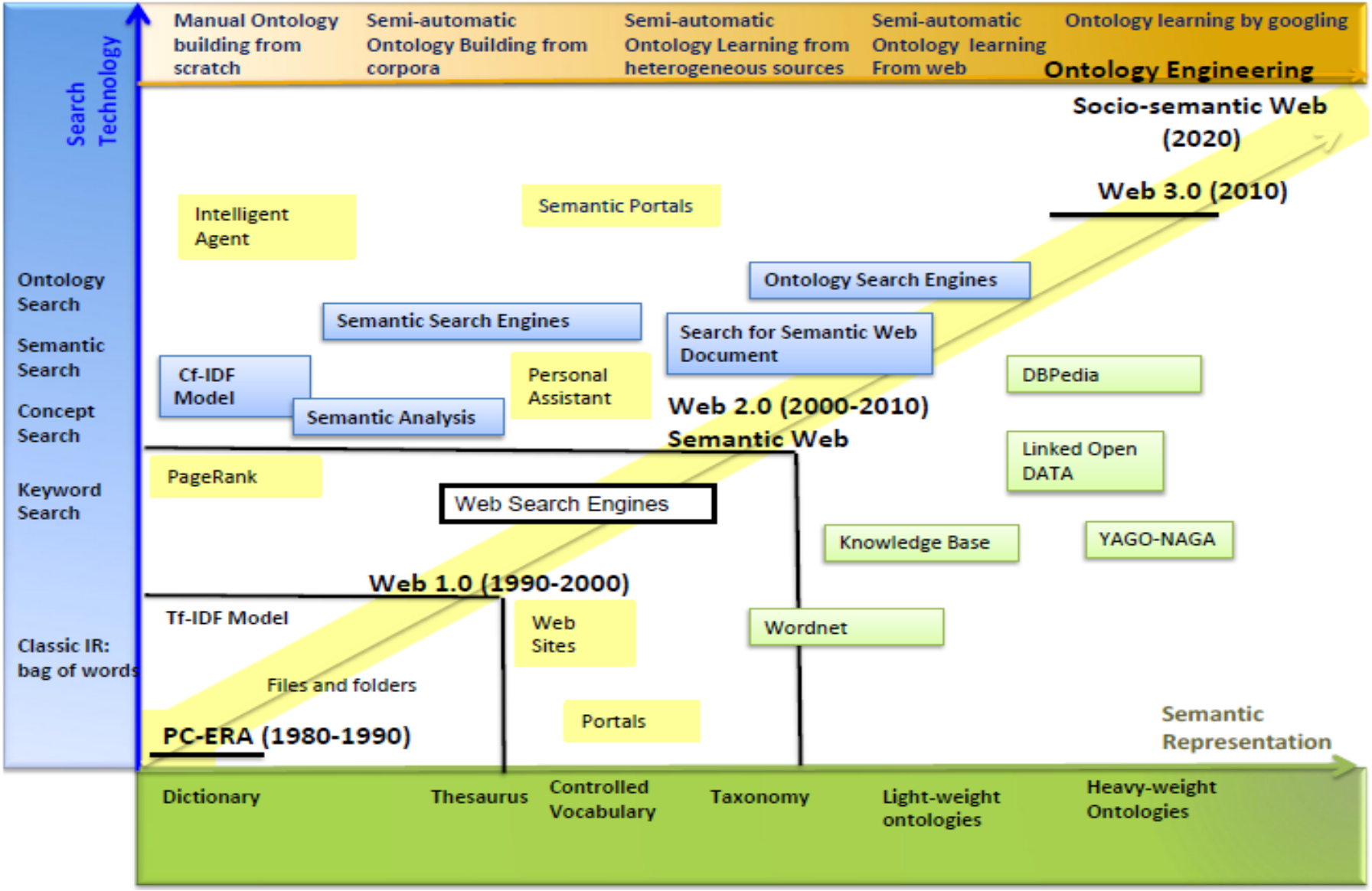
Unsupervised Web-based semantic relatedness for ontology learning : problems

- The use of statistical assessment is not sufficient to identify that a given concept candidate or relation is enough relevant to add it to the ontology

→ Using linguistic patterns to discover relations between concepts

$$\text{Score}_{\text{pattern}}(\text{candidate}) = \underset{i=1..n}{\overset{Patterns}{Max}}\left(\frac{hits(\text{" } pattern_i(\text{"}concept\text{"},\text{"candidate"})}{hits(\text{"candidate"})}\right)$$

Synthesis


Search Technology
Manual Ontology building from scratch
Semi-automatic Ontology Building from corpora
Semi-automatic Ontology Learning from heterogeneous sources
Semi-automatic Ontology learning From web
Ontology learning by googling
Ontology Engineering
Socio-semantic Web (2020)
Web 3.0 (2010)
Intelligent Agent
Semantic Portals
Ontology Search
Semantic Search
Concept Search
Keyword Search
Ontology Search Engines
Semantic Search Engines
Search for Semantic Web Document
Cf-IDF Model
Personal Assistant
DBPedia
Semantic Analysis
Web 2.0 (2000-2010)
Semantic Web
Linked Open DATA
PageRank
Web Search Engines
Knowledge Base
YAGO-NAGA
Web 1.0 (1990-2000)
Tf-IDF Model
Web Sites
Wordnet
Classic IR: bag of words
Files and folders
Portals
PC-ERA (1980-1990)
Semantic Representation
Dictionary
Thesaurus
Controlled Vocabulary
Taxonomy
Light-weight ontologies
Heavy-weight Ontologies

Ontology Leaning for Business Intelligence

Ontology Learning for Business Intelligence

- Engineering of Semantic Business Intelligence
  - Unified Knowledge representation
  - Semantic integration of multiple sources
  - Semantic search over heterogeneous data
- Semantic Technologies
  - NLP techniques for question-answering
  - Semantic annotations
  - Ontology-based Information Extraction

Ontology Leaning for Business Intelligence

Semantic Search

Main objectives

- PARLANCE: Probabilistic Adaptive Real-Time Learning And Natural Conversational Engine
- Goal: design and build mobile applications that approach human performance in conversational interaction, specifically in terms of the interactional *skills* needed to do so
- All of these *skills* will be learned or adapted using real data

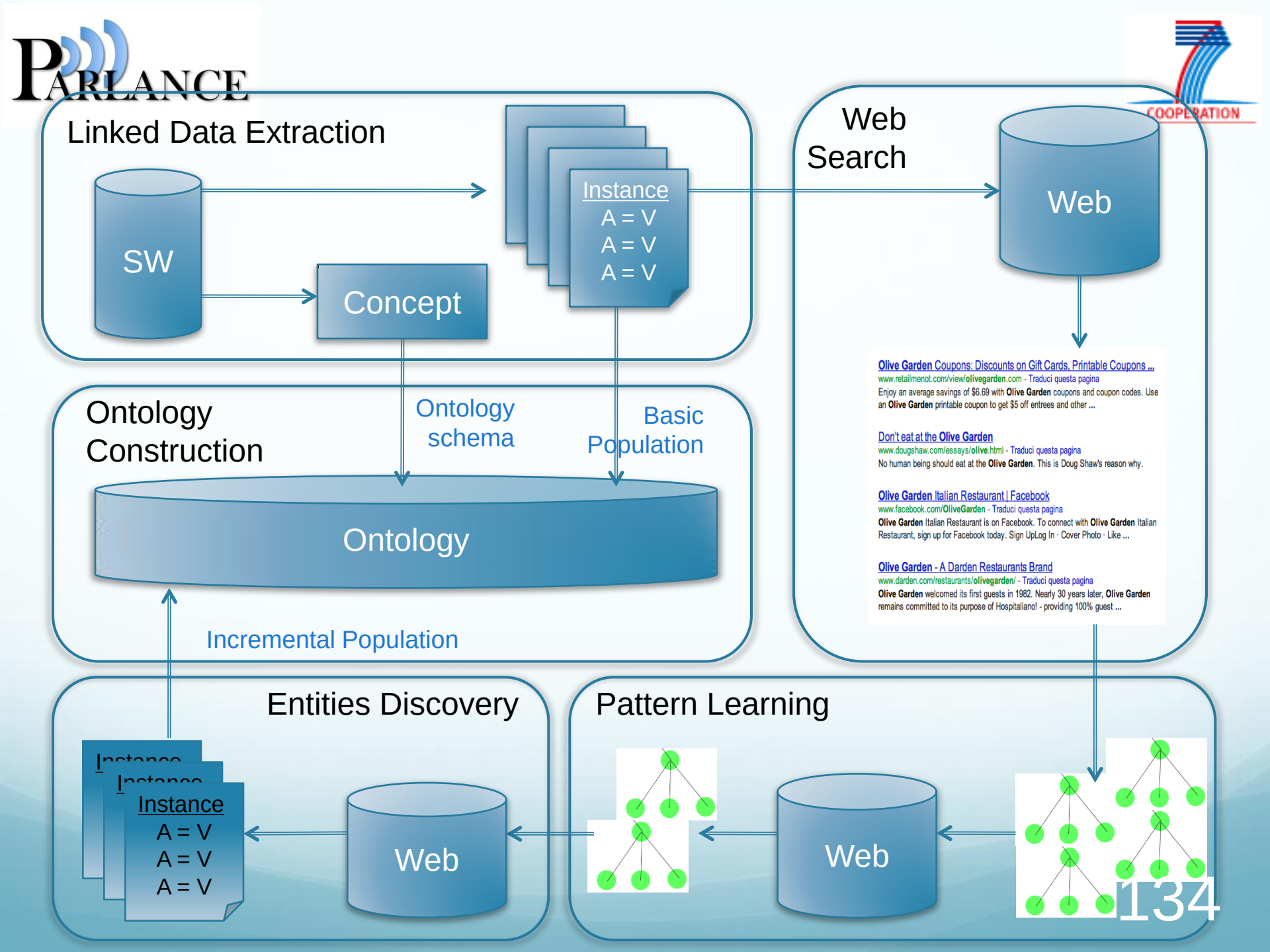
Main objectives

- Develop dialogue systems in 3 languages that are:
  - Incremental: dialogue act units
  - Personalized: adapt to different users with different goals in different contexts
  - Dynamic and evolving: possible to incorporate new concepts
  - Interactive hyper-local: take into account the location and surroundings of the user

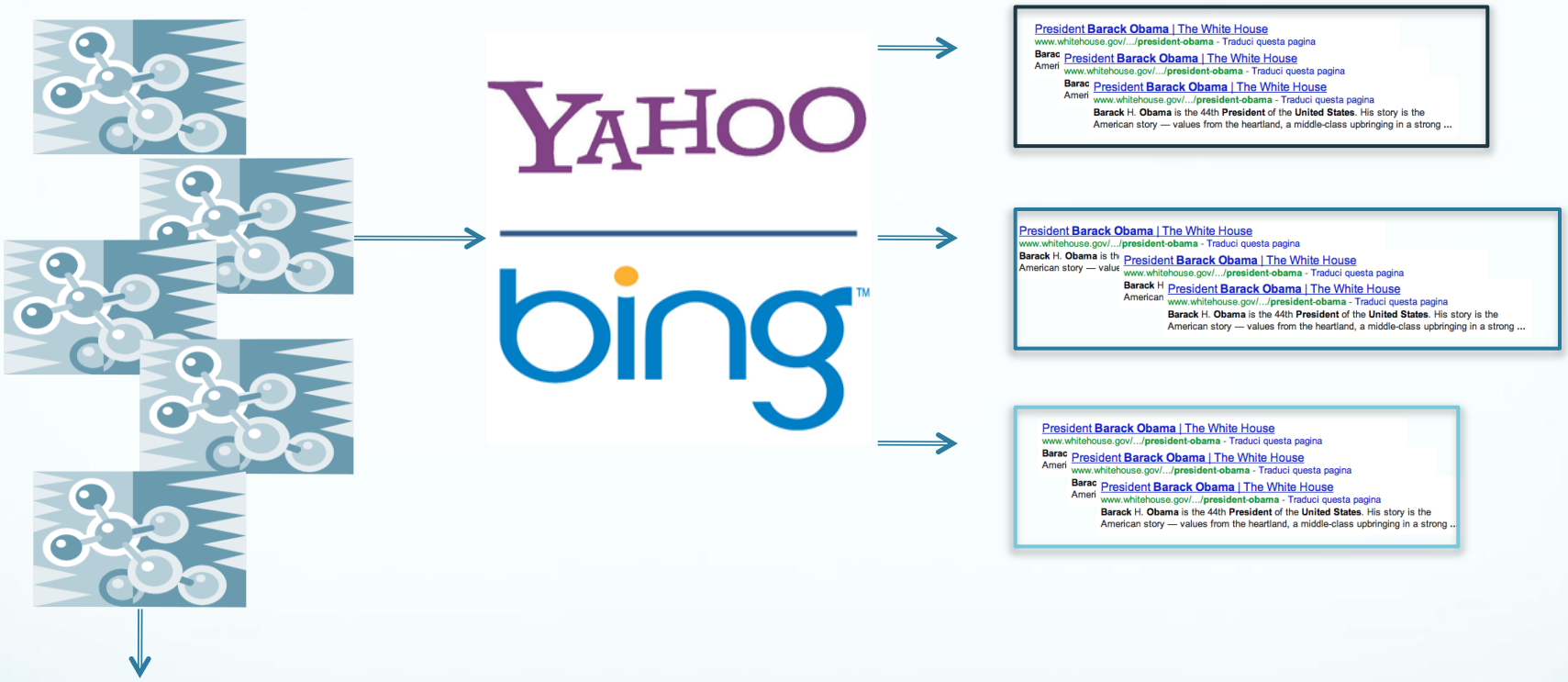
Our focus

- Modular incremental ontology enrichment
  - Using existing knowledge bases
  - Ontology learning from Web
  - Query-driven knowledge base enrichment

- Dynamic, evolving rich user models
  - Static information
  - Contextual information
  - Social information – collaborative filtering


Linked Data Extraction
SW
Concept
Instance
A = V
A = V
A = V
Web Search
Web
Olive Garden Coupons: Discounts on Gift Cards, Printable Coupons ...
www.retailmenot.com/view/olivegarden.com - Traduci questa pagina
Enjoy an average savings of $6.69 with Olive Garden coupons and coupon codes. Use an Olive Garden printable coupon to get $5 off entrees and other ...
Don't eat at the Olive Garden
www.dougshaw.com/essays/olive.html - Traduci questa pagina
No human being should eat at the Olive Garden. This is Doug Shaw's reason why.
Olive Garden Italian Restaurant | Facebook
www.facebook.com/OliveGarden - Traduci questa pagina
Olive Garden Italian Restaurant is on Facebook. To connect with Olive Garden Italian Restaurant, sign up for Facebook today. Sign UpLog In · Cover Photo · Like ...
Olive Garden - A Darden Restaurants Brand
www.darden.com/restaurants/olivegarden/ - Traduci questa pagina
Olive Garden welcomed its first guests in 1982. Nearly 30 years later, Olive Garden remains committed to its purpose of Hospitaliano! - providing 100% guest ...
Ontology Construction
Ontology schema
Basic Population
Ontology
Incremental Population
Entities Discovery
Instance
A = V
A = V
A = V
Web
Pattern Learning
Web

Web Search, Web Corpus



"Barack Obama + United States"

President **Barack Obama** | The White House

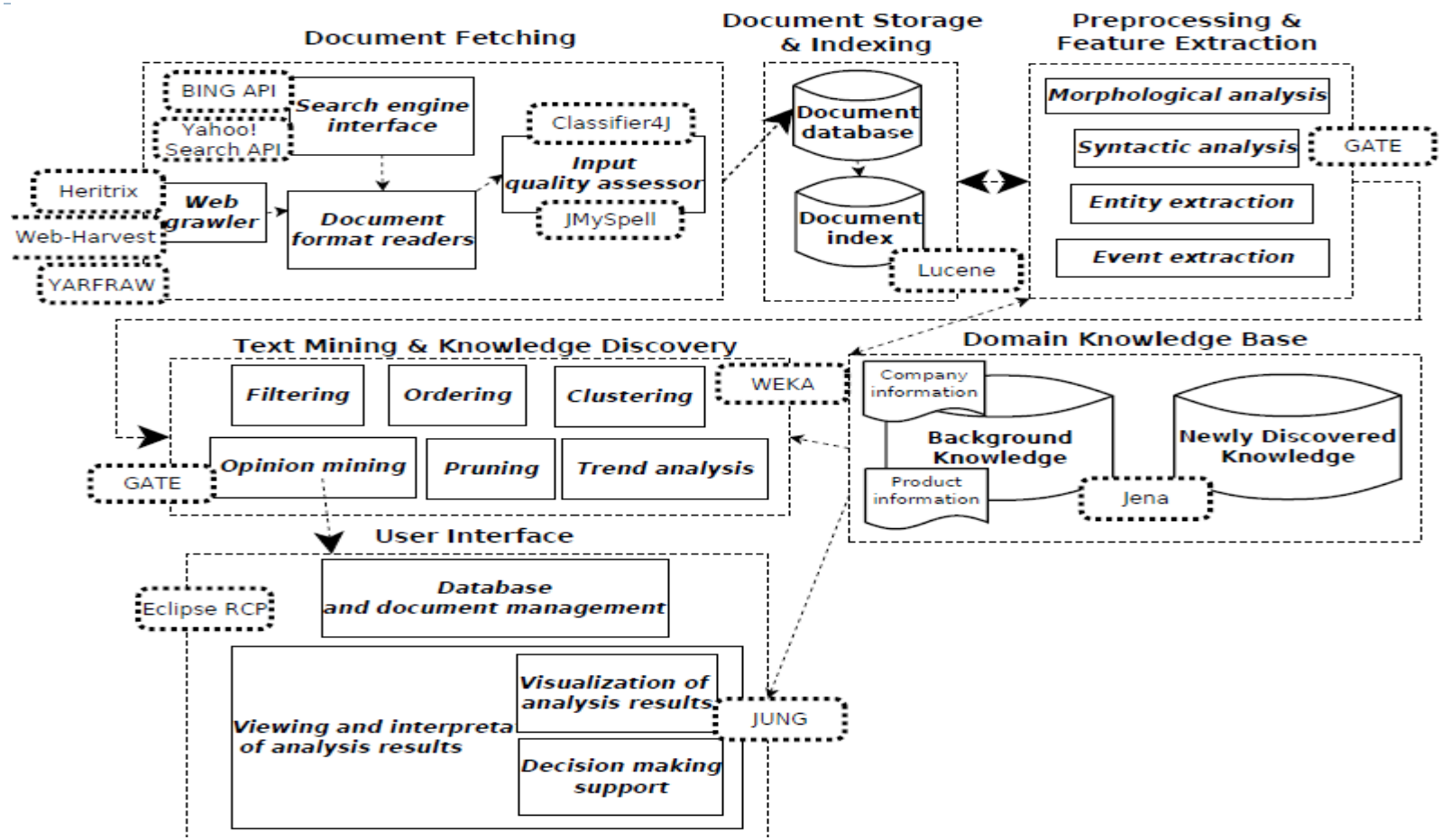
www.whitehouse.gov/.../**president-obama** - Traduci questa pagina

Barack H. **Obama** is the 44th **President** of the **United States**. His story is the American story — values from the heartland, a middle-class upbringing in a strong ...

Ontology Leaning for Business Intelligence

Ontology-based Information Extraction

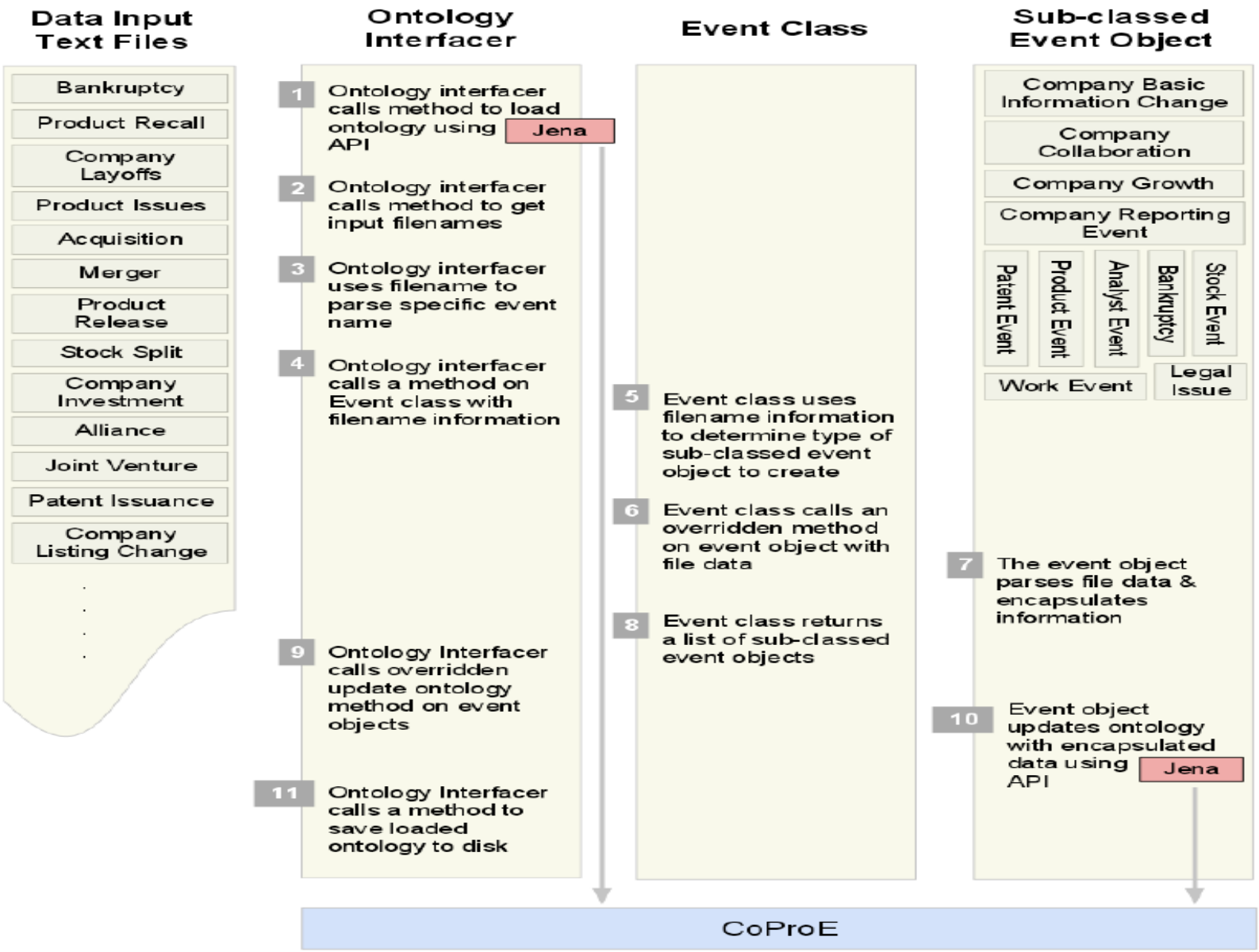
Ontologies for Collecting and Analyzing Business Intelligence
DAVID (Data Analysis and Visualization aID) system



Kakkonen, T., Mufti, T.: Developing and Applying a Company, Product and Business Event Ontology for T Mining. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, Graz, Austria, 2011

Ontology Learning for Business Intelligence

Ontologies for Collecting and Analyzing Business Intelligence



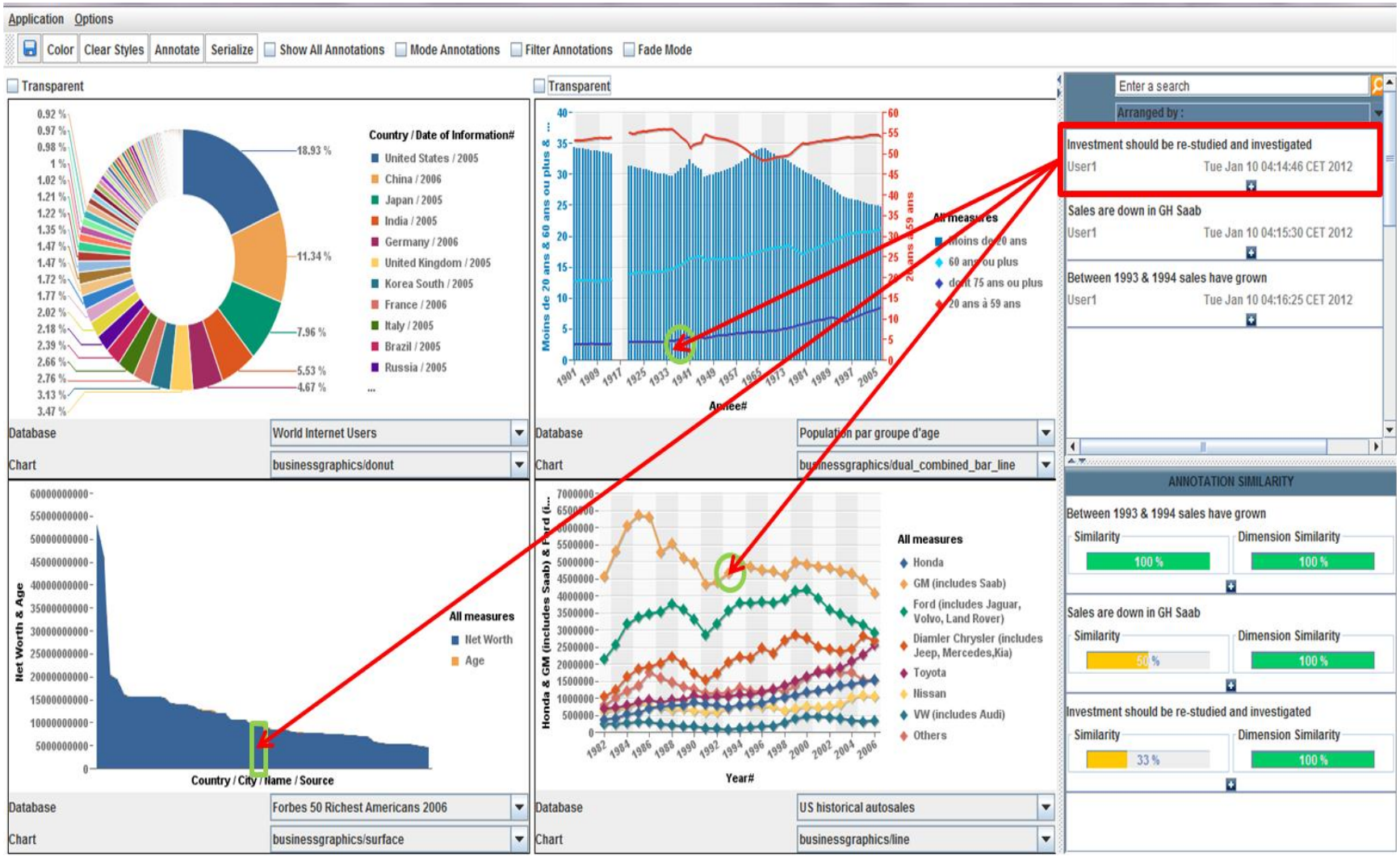
Ontology Leaning for Business Intelligence

Annotation of Dashboard

Annotating BI Visualization Dashboards: Needs & Challenges. Micheline Elias and Anastasia Bezerianos

ACM SIGCHI Conference on Human Factors in Computing Systems, CHI 2012.

Ontological Need:
Enriching Ontologies From annotations ??



Conclusion

Conclusion

- Bridging the gap between academic research approaches and real business use cases
- Scalability of knowledge
  - Modular ontologies
  - Modular reasoning
- Open Linked data and Privacy issues

Toward Open Linked Business Processes ?!

Questions??