BRIDGING THE WEB OF DOCUMENTS WITH THE WEB OF DATA AT *SEARCH TIME*

Yannis Tzitzikas

University of Crete and

FORTH-ICS

FORTH Institute of Computer Science

ACKNOWLEDGEMENTS

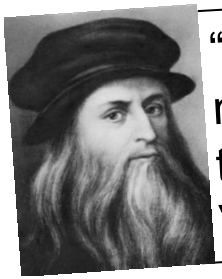
- Research done mainly with my students (more in the acknowledgements)
  - Pavlos Fafalios

  - Panagiotis Papadakos

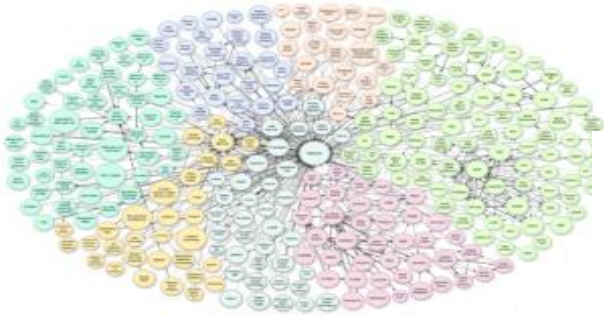
Outline

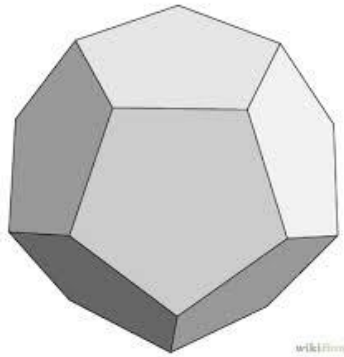
- Introduction and Background (60')
  - Exploratory Search
  - Semantic Web and Linked Data
  - Faceted Exploration
- Bridging the Web of Documents with the Web of Data (50')
  - Possible Ways
  - Focus on doing this at search time
  - Case studies: presented as a series of 9 milestones
- Synopsis, Challenges, Discussion (10')

- References
- Acknowledgements



"Study without desire spoils the memory, and it retains nothing that it takes in." Leonardo da Vinci





1. Introduction and Background

Exploratory Search

Semantic Web and Linked Data

Faceted Exploration

What Users usually Want/Do when Searching?

Kinds of information needs

- *Precision-oriented*
  - Locate one resource or/and its attributes

    e.g. Find the telephone of a store
- *Recall-oriented*
  - Locate a **set** of resources

    e.g. Medical information seeking, travel planning

Over 60% of web search queries are ***recall-oriented***

[Broder 02, Rose and Levinson 04]

Recall-Oriented Information Needs

- In Recall-Oriented Information Needs:
  - the users require >1 hit
  - essentially such needs correspond to decision tasks
- Examples of Recall-oriented information needs
  - Booking
  - Product-buying
  - Bibliography search
  - Patent Search
  - Medical Search
  - ….

QUESTIONS

How many of you (raise your hand):

- have bought your mobile phone by looking at the first hit(s) that google returned to your query?
- have ever booked the 1st hotel returned by google?
- have entirely read the first paper that Scholar Google returned to your query?
- have made at least once a search about a medical issue?
  - keep your hand raised if you had submitted to Google only one query
    - keep your hand raised if you have read only the first 5 hits of that single query that you submitted to Google, and eventually that was enough for fully satisfying your information need

EXPLORATORY SEARCH

Wikipedia:

"Exploratory search is a specialization of information exploration which represents the **activities carried out by searchers** who are either:

a) unfamiliar with the domain of their goal (i.e. need to learn about the topic in order to understand how to achieve their goal)

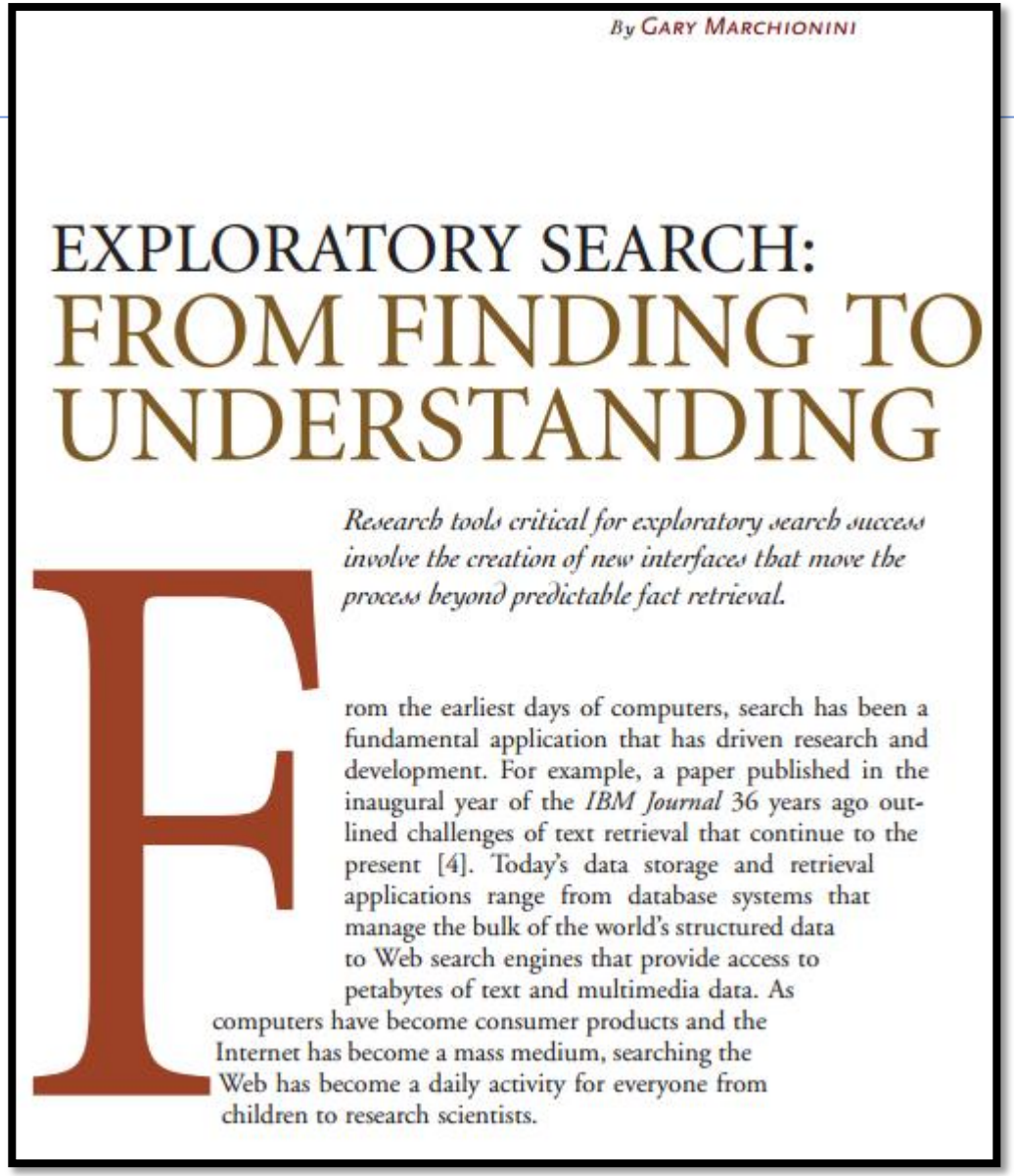
b) unsure about the ways to achieve their goals (either the technology or the process)

c) or even unsure about their goals in the first place.

Consequently, exploratory search covers **a broader class of activities** than typical information retrieval, such as ***investigating***, ***evaluating***, ***comparing***, *and* ***synthesizing***, where new information is sought in a defined conceptual area; exploratory data analysis is another example of an information exploration activity. Typically, therefore, such users generally **combine querying and browsing** strategies to foster learning and investigation."

THEREFORE…

- Ranking is not enough for exploratory search



By GARY MARCHIONINI

EXPLORATORY SEARCH: FROM FINDING TO UNDERSTANDING

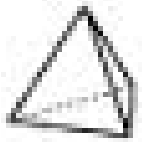
Research tools critical for exploratory search success involve the creation of new interfaces that move the process beyond predictable fact retrieval.

From the earliest days of computers, search has been a fundamental application that has driven research and development. For example, a paper published in the inaugural year of the *IBM Journal* 36 years ago outlined challenges of text retrieval that continue to the present [4]. Today's data storage and retrieval applications range from database systems that manage the bulk of the world's structured data to Web search engines that provide access to petabytes of text and multimedia data. As computers have become consumer products and the Internet has become a mass medium, searching the Web has become a daily activity for everyone from children to research scientists.

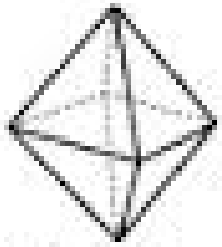
Some Common Requirements for Effective Exploratory Search

- Allow easy and fast access even to **low ranked** hits
- Allow browsing and inspecting the found hits in **groups** (according to various criteria)
- Offer **overviews** of the search results
  - Compute and show descriptions and **count** information for the various groups, or other **aggregated** values
- Allow **gradual** restriction/ranking of the search results

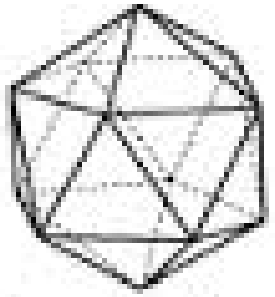
FACETED SEARCH



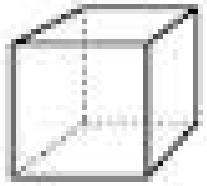
Tetrahedron



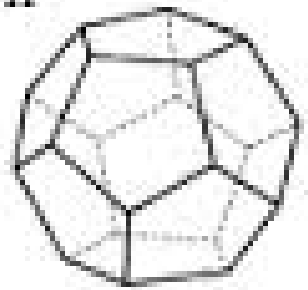
Octahedron



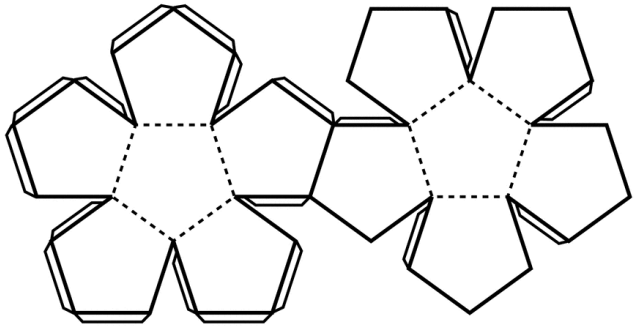
Icosahedron



Cube



Dodecahedro



FACETED SEARCH/EXPLORATION

Faceted Exploration is a **widely** used interaction scheme **for Exploratory Search**

A short (and rather informal) definition:

FE is a **session-based** interactive method for **query formulation** (commonly over a multidimentional information space) through simple clicks offering

- ✓ an overview of the result set (groups and count information)
- ✓ never leading to empty results sets

- The access paradigm supported is a *conceptual exploration*,
  - far more frequent in "search" tasks than the retrieval by exact specification supported by search engines and database queries.
- Simple and easily understood by users.

FACETED SEARCH

Wikipedia:

- **Faceted search**, also called **faceted navigation** or **faceted browsing**, is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, called facets, enabling the classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined, taxonomic order.
- Facets correspond to properties of the information elements. They are often derived by analysis of the text of an item using entity extraction techniques or from pre-existing fields in a database such as author, descriptor, language, and format. Thus, existing web-pages, product descriptions or online collections of articles can be augmented with navigational facets.
- Within the academic community, faceted search has attracted interest primarily among library and information science researchers, and to some extent among computer science researchers specializing in information retrieval

A part of My incomplete facebook of faceted Classification/Browsing/search

S. R. Ranganathan
(1892-1972
Faceted
Classification

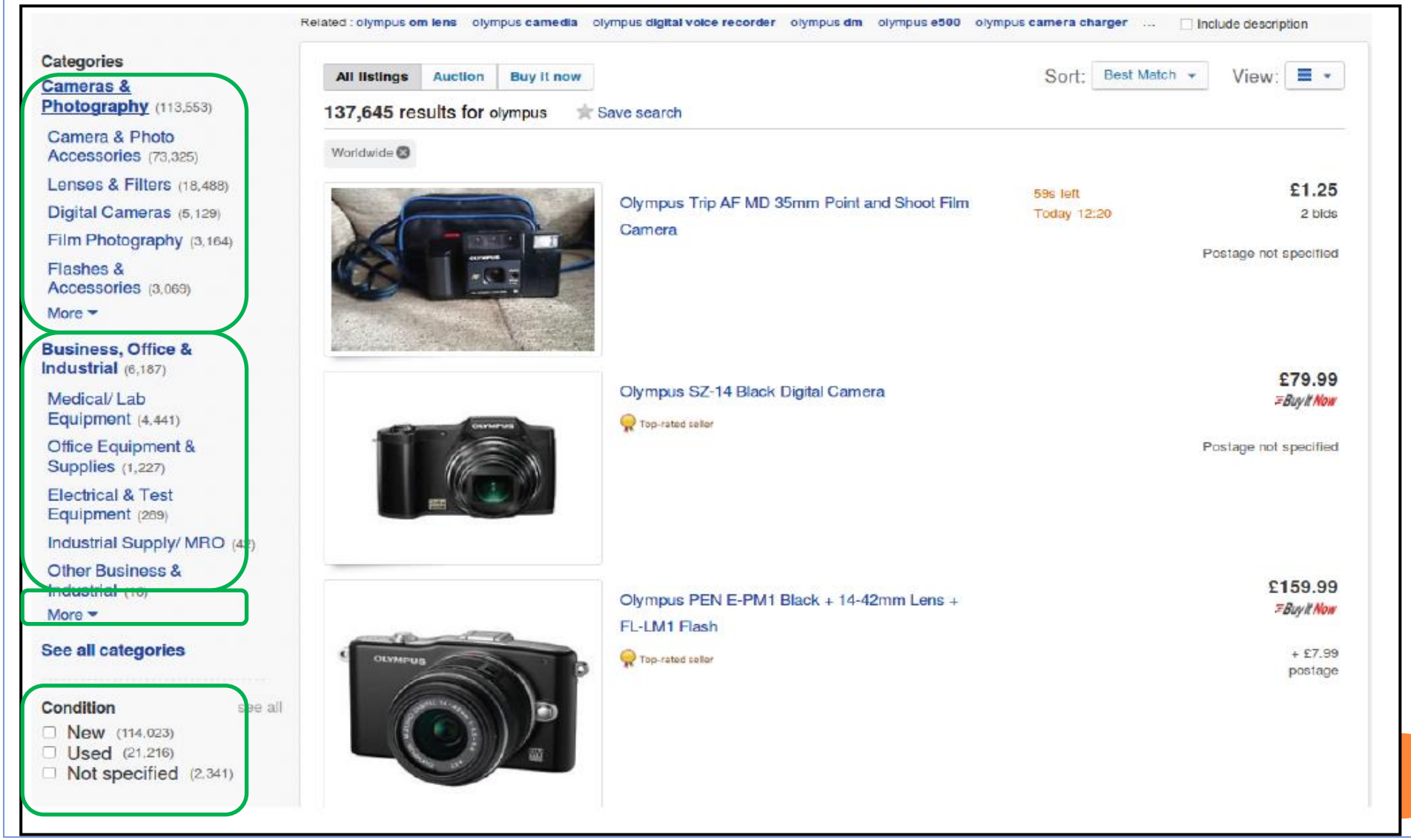
Giovanni Maria
Sacco
Dynamic Taxnomies

Sébastien Ferré
University Rennes 1

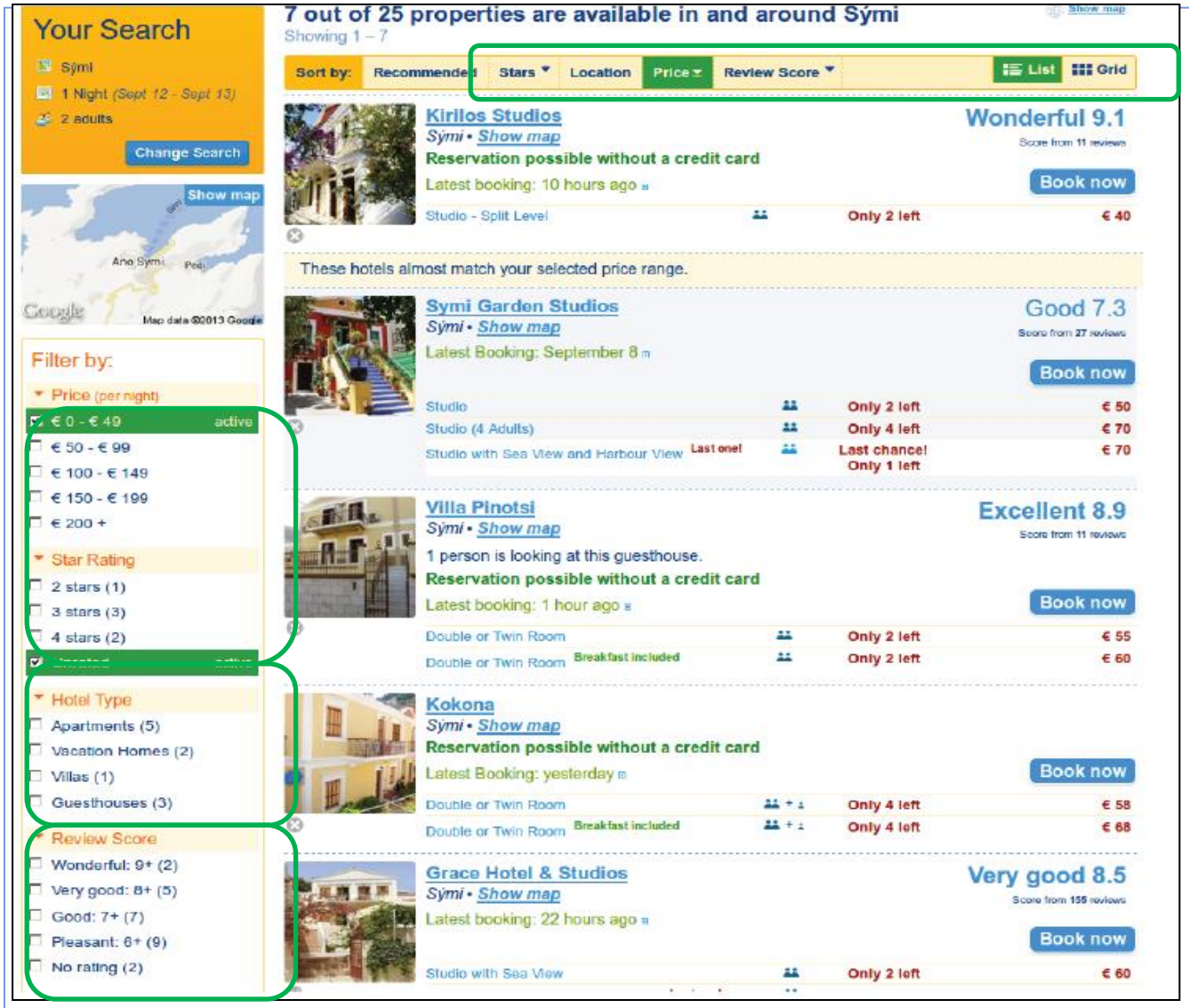
Marti A. Hearst
University of
California, Berkeley

- Let's now see some examples from some widely used systems

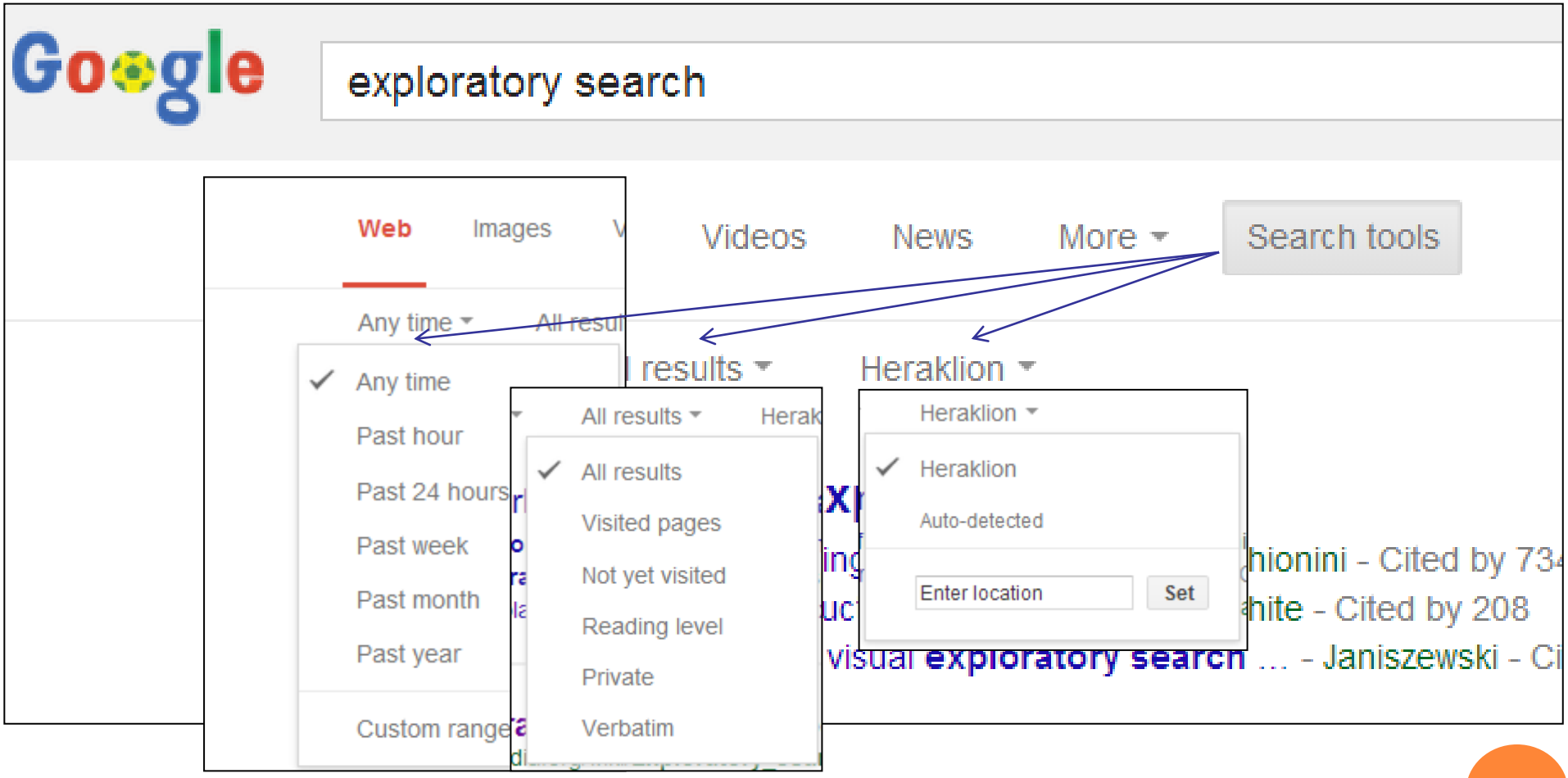
EXAMPLE: EBAY



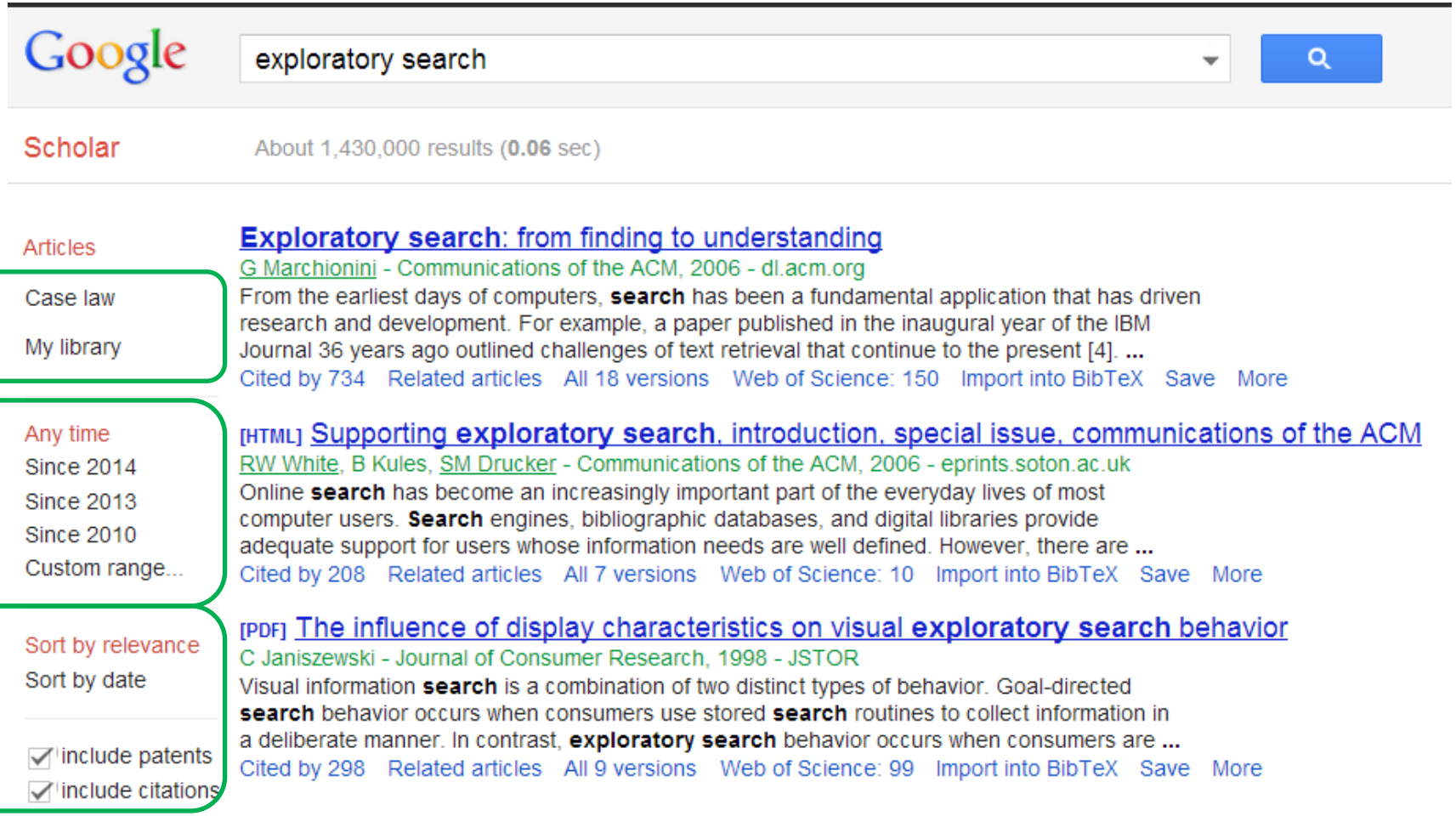
EXAMPLE OF FDT: BOOKING.COM



EXAMPLE: GOOGLE SEARCH (LIMITED FUNCTIONALITY: NO COUNT INFORMATION)



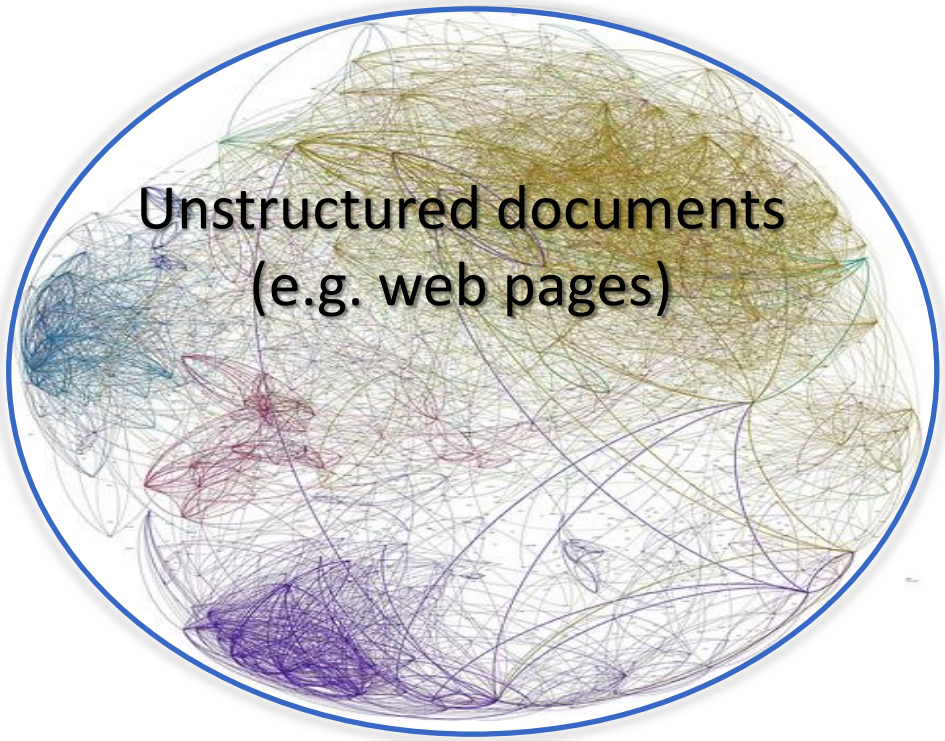
EXAMPLE: SCHOLAR GOOGLE

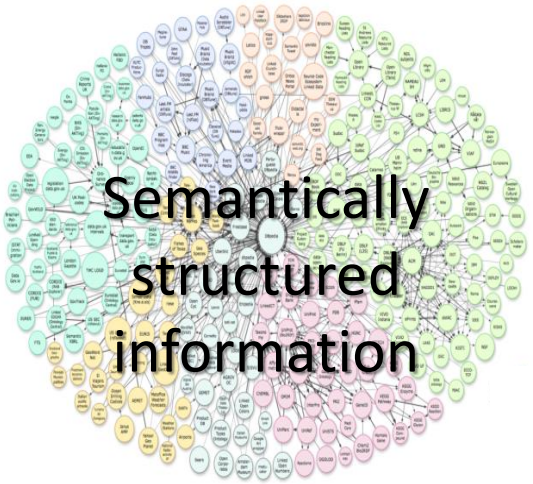


More on FE

- We will return to define more formally the interaction of faceted exploration, after first making a short introduction of the Semantic Web and Linked Data.
  - After that we will show how FE can be applied on such data

Relatively Recently
a … second "world" starts growing





Semantic Web and Linked Data

The Semantic Web Vision

- The Semantic Web is an evolving extension of the WWW where the **content** can be expressed not only in natural language but also in formal languages (e.g. RDF/S, OWL) that can be read and used by software agents, permitting them to find, share and integrate information more easily

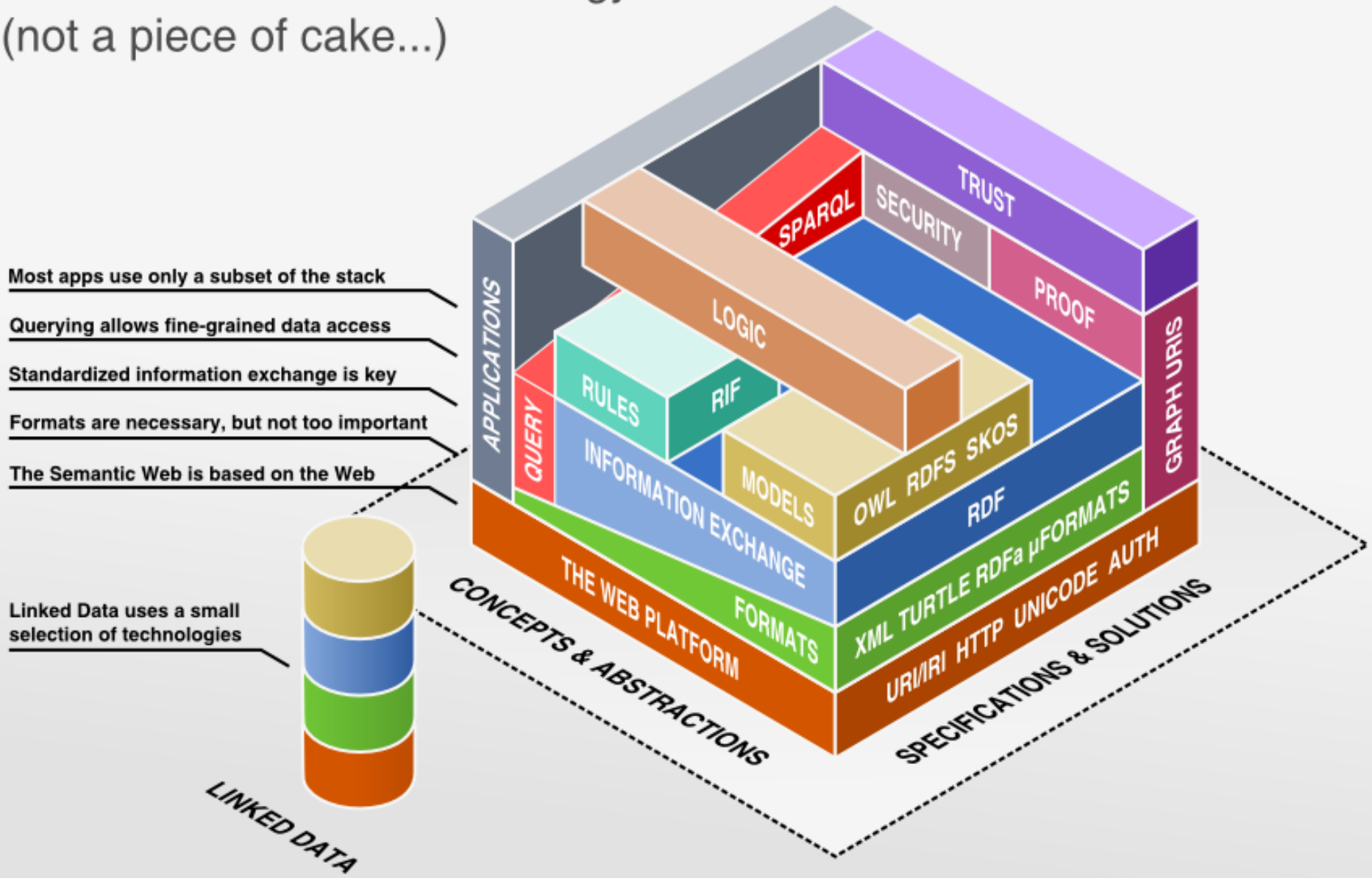
- *Imagine that the objective is the collaborative creation and evolution of a world wide distributed graph (about everything ☺)*
  - *We could say that this graph resembles the structure of an EntityRelationship Diagram (recall Databases)*

SEMANTIC WEB TECHNOLOGIES

- For achieving the Semantic Web vision, the recent years several technologies have been emerged, many of them are international (W3C) standards

- These technologies include:
  - knowledge representation languages (e.g. RDF/S, OWL) and formats for exchanging knowledge
  - query languages (π.χ. SPARQL),
  - rule languages and inference engines
  - Techniques for constructing mappings for integrating/harmonizing schemas and data
  - Technologies for mining structure knowledge from texts
  - Various APIs.
- Linked Open Data Are based on these technologies.

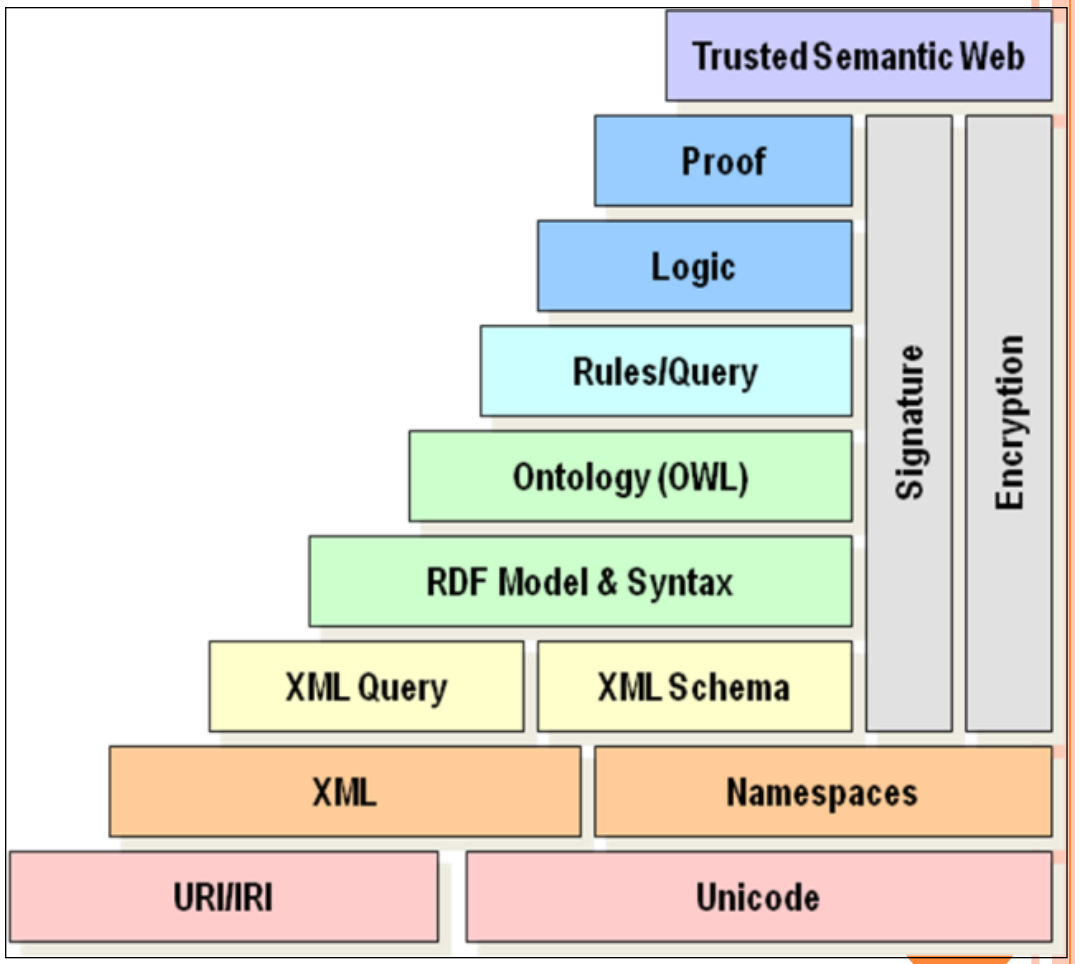
The Semantic Web Technology Stack
(not a piece of cake...)



THE TECHNOLOGY STACK OF SEMANTIC WEB

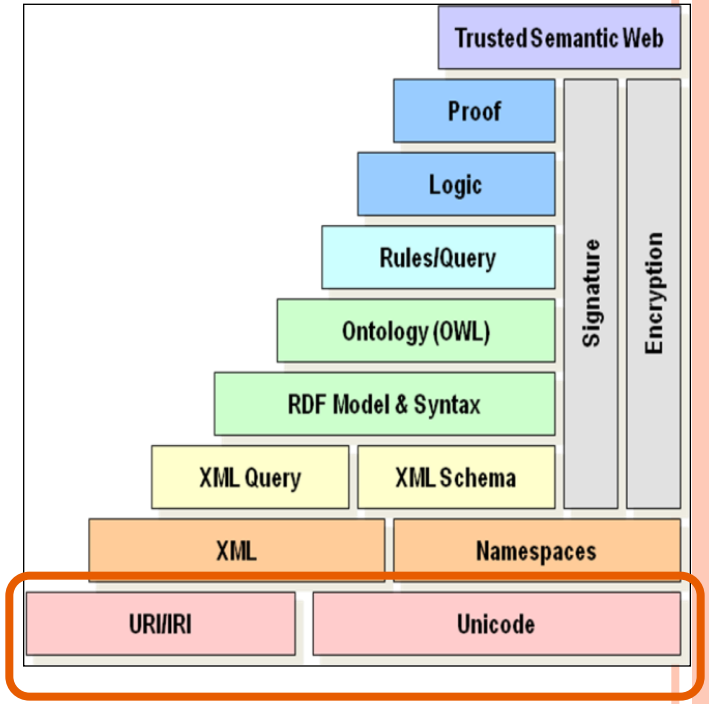
By starting from the World Wide Web where the content is represented as hypertext (i.e. pages containing texts and links to other pages), let's now give an overview of the basic technologies on which the Semantic Web is based.

The current technology stack of Semantic Web can be illustrated as follows:



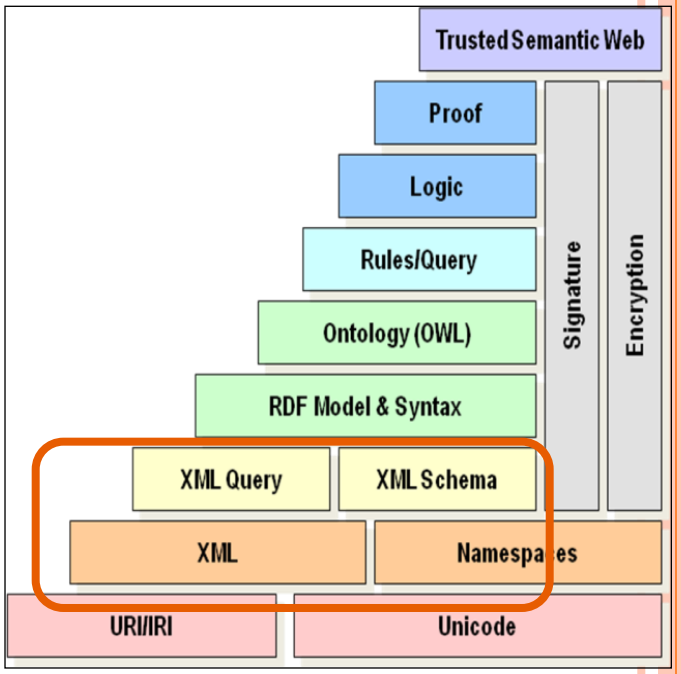
TECHNOLOGICAL STACK: URIs & UNICODE

- At the lowest level we have the URI (Uniform Resource Identifiers) for identifying and locating resources, and UNICODE for representing characters from various natural languages



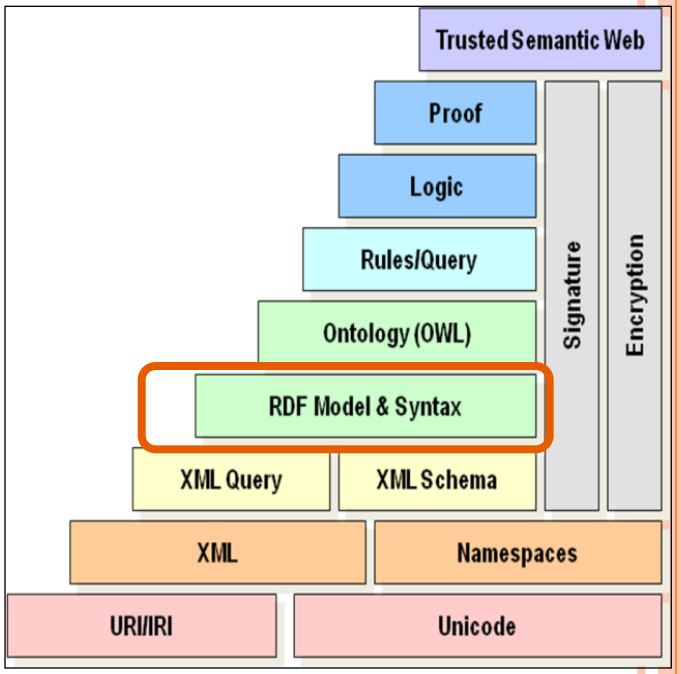
TECHNOLOGICAL STACK: XML

- XML provides syntax for having structured documents, but does not impose any constraint regarding the semantics of these documents.
- XML Schema offers a method to restrict the structure that XML documents can have.
- For querying XML documents, we have XPath and XQuery.



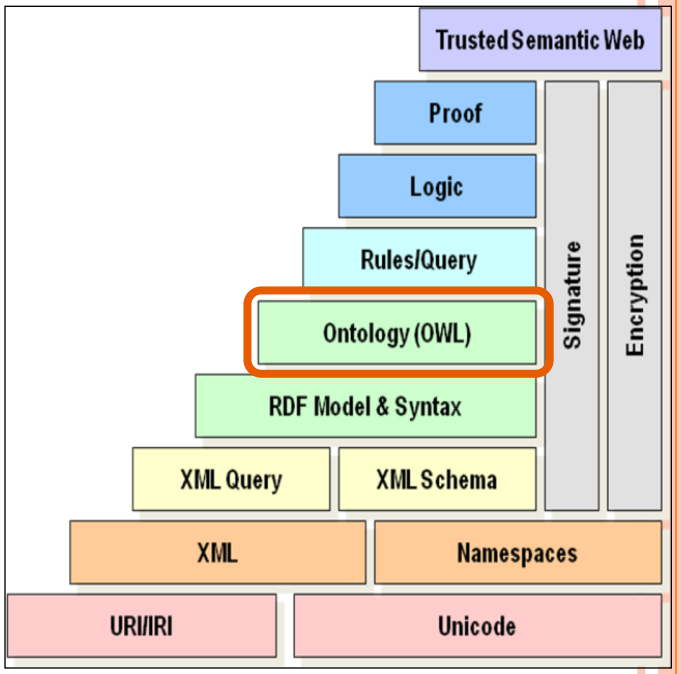
TECHNOLOGICAL STACK: RDF & RDFS

- RDF (Resource Description Framework) is a structurally object-oriented model for representing objects (resources) and associations between them.
- It allows expressing content in form of triples (*subject, predicate, object*), and sets of triples actually form a semantic network/graph.
- These triples can be expressed in various formats (TriG, N3 RDF/XML), some of them are based on XML (RDF/XML)
- RDFS allows defining the vocabulary to be used in RDF and semantic relationships between the elements of this vocabulary



TECHNOLOGICAL STACK: OWL

- The Web Ontology Language (OWL) is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases. The languages are characterized by formal semantics and RDF/XML-based serializations for the Semantic Web.

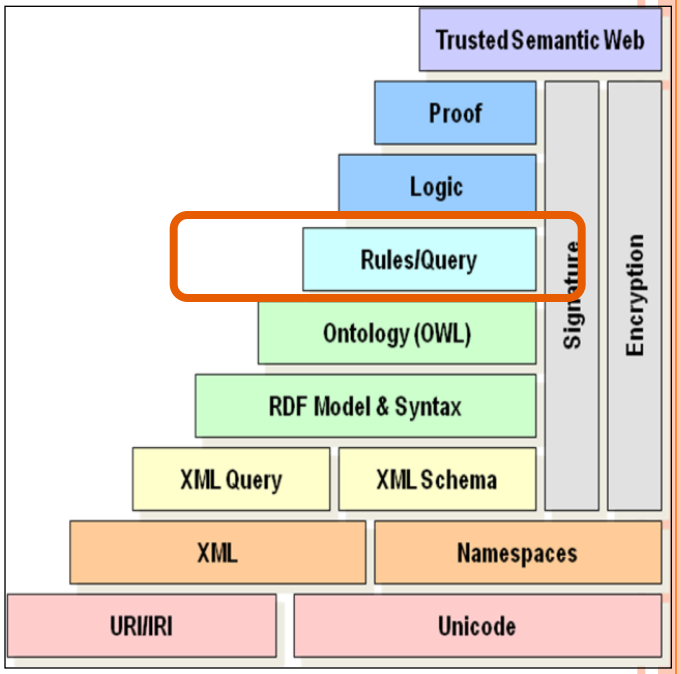


TECHNOLOGICAL STACK: SPARQL & SWRL

- For exploiting the structured content that has been represented using RDF/S, there are query languages and rule languages

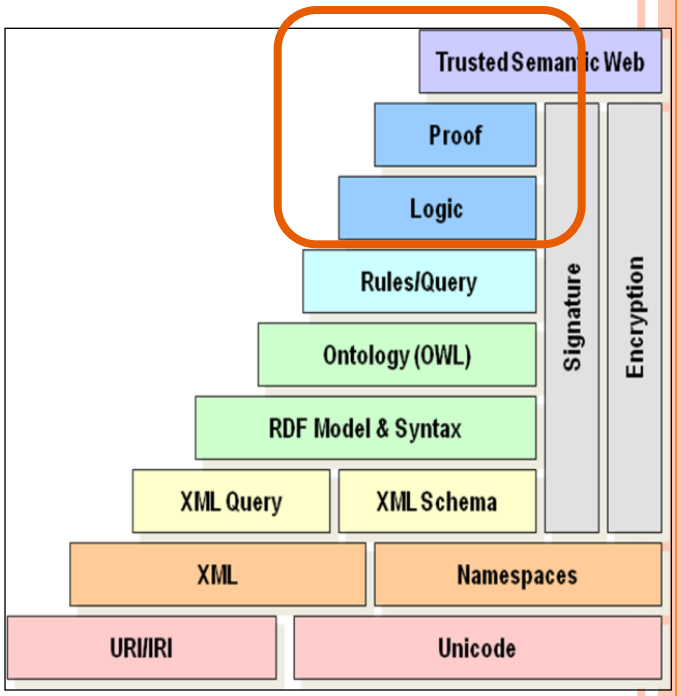
- Specifically, SPARQL (SPARQL Protocol and RDF Query Language) is a query language for knowledge expressed in RDF/OWL.

- SWRL (Semantic Web Rule Language) allows expressing inference rules (essentially Horn rules).



TECHNOLOGICAL STACK: LOGIC, PROOF, TRUST

- The layers *Logic* and *Proof* concern the enrichment of the expressiveness of the representation languages
- Finally the *Trust* layer concerns trust issues, e.g. digital signatures (for proving that one particular person has written or agrees with a particular document or sentence, as well as trust networks allowing users to define who they trust (and so on), eventually yielding trust networks (Web of Trust)

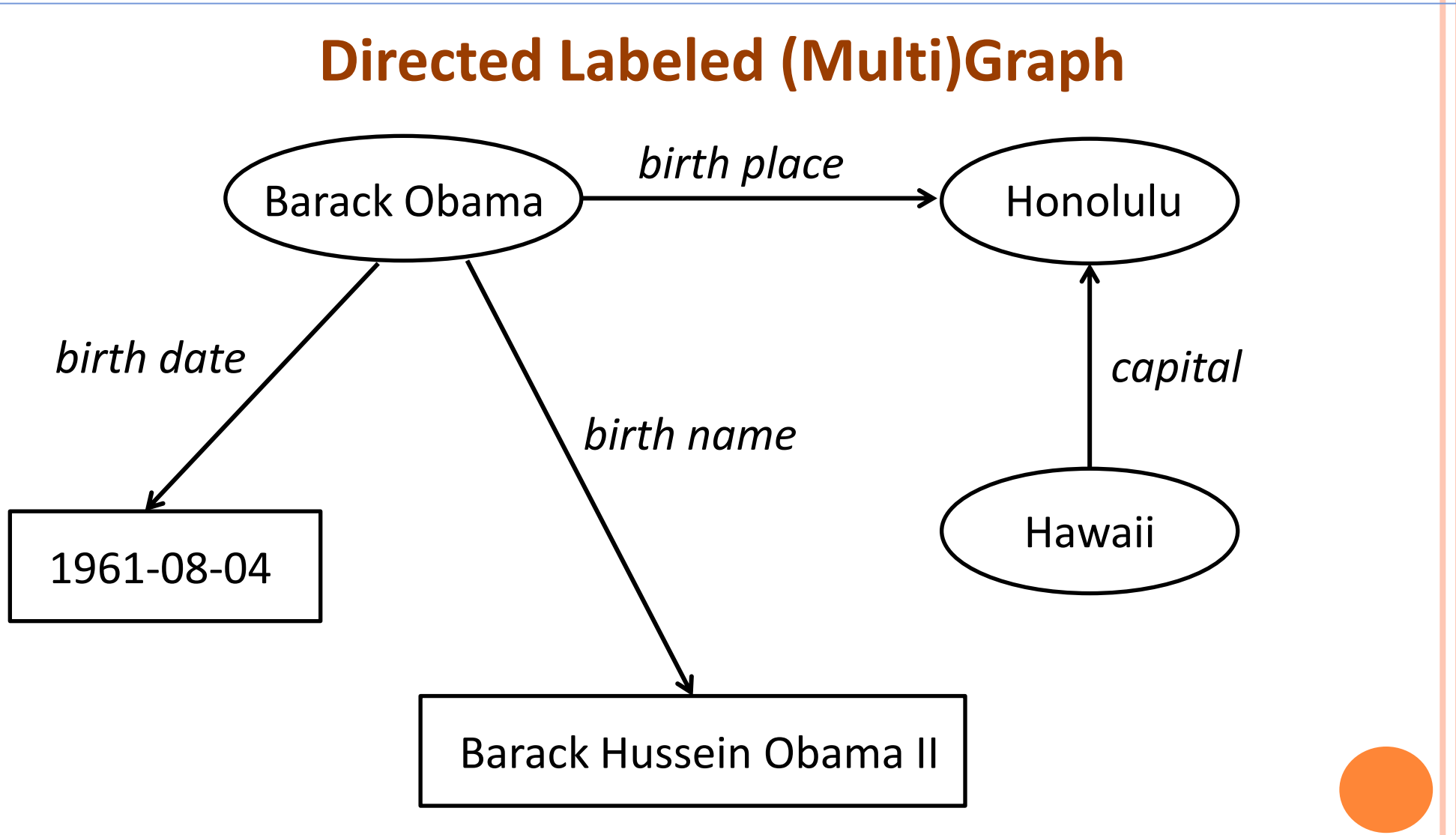


STANDARDIZATION, PERSISTENT STORAGE

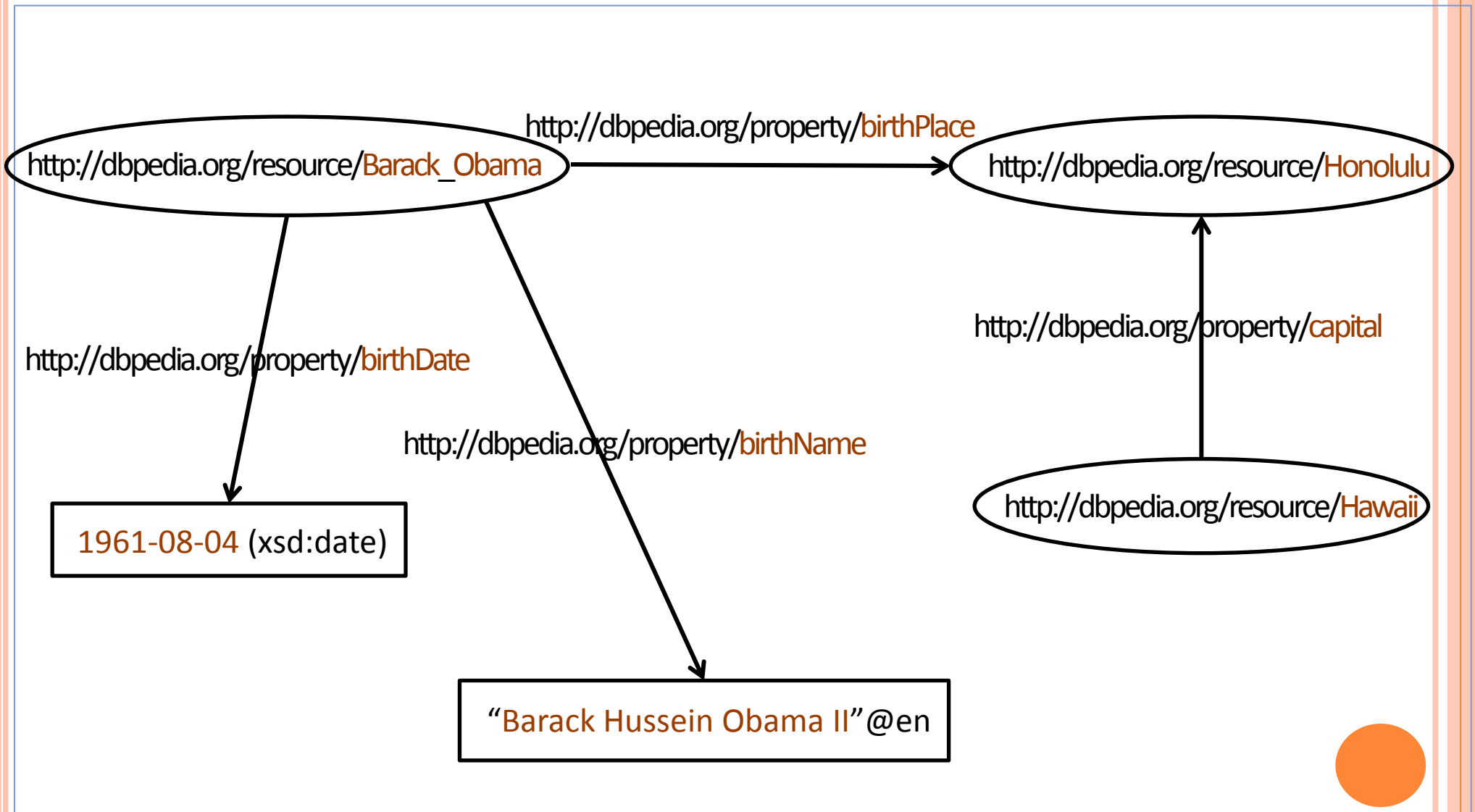
- The part of the stack that has been implemented and standardized allows describing resources (of digital objects mainly) using classes and properties that have been defined in ontologies (expressed using RDF/S or OWL) which are accessible from the network which in turn can be connected (one ontology can extend classes and properties of other ontologies).

- For the persistent storage of such contents (independently of whether it corresponds to data or metadata) in the form of connected semantic network (semantic graph), there are tools, usually referred to as triplestores. We could call them *semantic data bases* since they receive as input RDF/S data, support integrity constraints, enable querying (SPARQL support), and concurrent access.

Example of an RDF Graph



RDF Graph


http://dbpedia.org/property/birthPlace
http://dbpedia.org/resource/Barack_Obama
http://dbpedia.org/resource/Honolulu
http://dbpedia.org/property/capital
http://dbpedia.org/property/birthDate
http://dbpedia.org/property/birthName
http://dbpedia.org/resource/Hawaii
1961-08-04 (xsd:date)
"Barack Hussein Obama II"@en

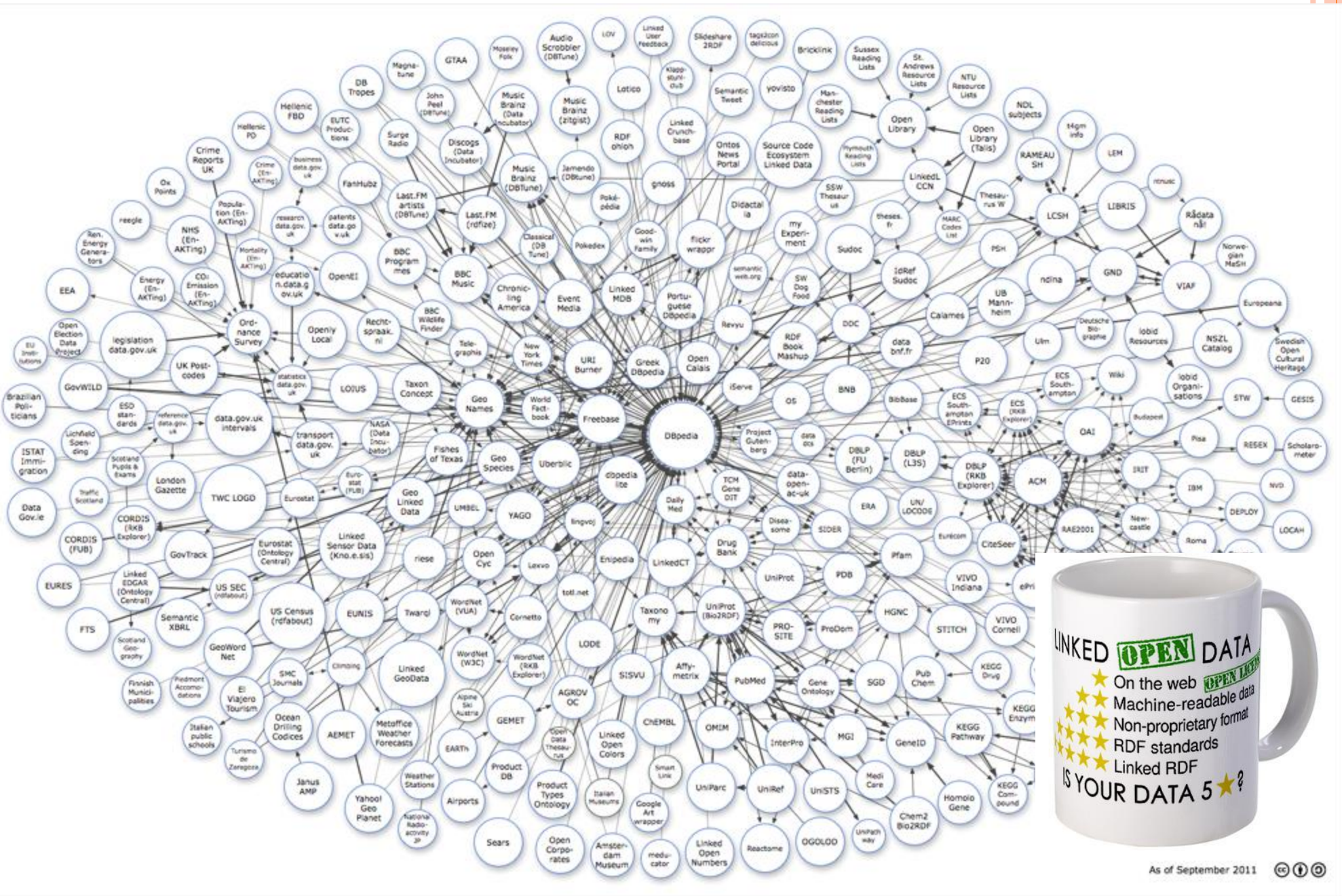
RDF Graph Representation

In XML (RDF/XML):

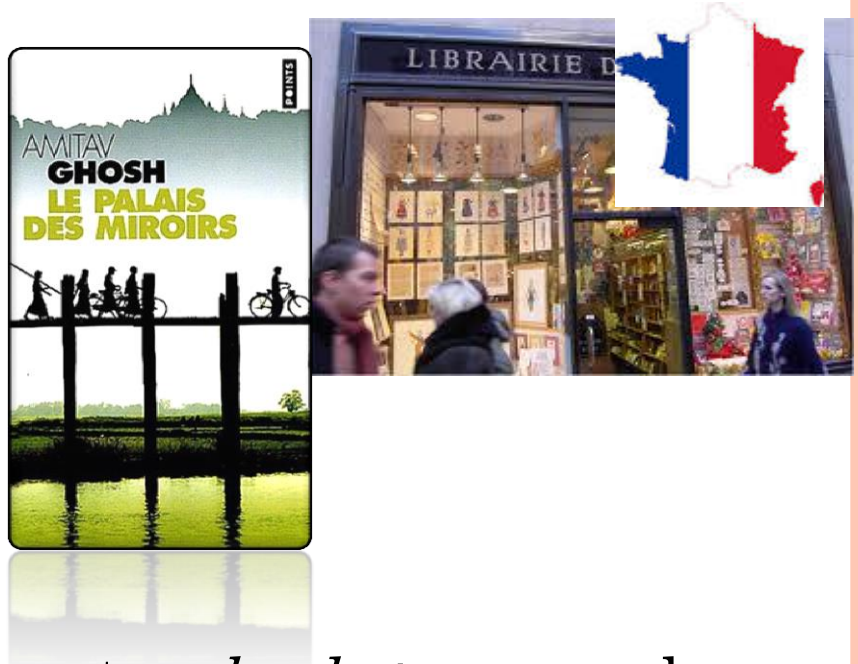
```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dbpprop="http://dbpedia.org/property/"
  <rdf:Description rdf:about="http://dbpedia.org/resource/Barack_Obama">
    <dbpprop:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
        1961-08-04
    </dbpprop:birthDate>
    <dbpprop:birthPlace rdf:resource="http://dbpedia.org/resource/Honolulu" />
    <dbpprop:birthName xml:lang="en">
        Barack Hussein Obama II
    </dbpprop:birthName>
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Hawaii">
    <dbpprop:capital rdf:resource="http://dbpedia.org/resource/Honolulu" />
  </rdf:Description>
</rdf:RDF>
```

In N-Triples:

```
<http://dbpedia.org/resource/Barack_Obama>  <http://dbpedia.org/property/birthPlace>    <http://dbpedia.org/resource/Honolulu> .
<http://dbpedia.org/resource/Barack_Obama>  <http://dbpedia.org/property/birthDate> "1961-08-04"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Barack_Obama>  <http://dbpedia.org/property/birthName> "Barack Hussein Obama II"@en .
<http://dbpedia.org/resource/Hawaii>    <http://dbpedia.org/property/capital>   <http://dbpedia.org/resource/Honolulu> .
```


DBpedia
Freebase
Geo Names
Linked GeoData
Linked Sensor Data (Kno.e.sis)
Linked Open Numbers
Linked Open Colors
Linked Open Corporates
Linked Open Amsterdam Museum
Linked Open Reactome
Linked Open OGOLOD
Linked Open Chem2 Bio2RDF
Linked Open Homolo Gene
Linked Open KEGG Compound
Linked Open KEGG Pathway
Linked Open KEGG Enzyme
Linked Open KEGG Drug
Linked Open Pub Chem
Linked Open GeneID
Linked Open MGI
Linked Open InterPro
Linked Open UniSTS
Linked Open UniRef
Linked Open UniParc
Linked Open UniProt (Bio2RDF)
Linked Open UniProt
Linked Open Taxonomy
Linked Open PubMed
Linked Open Gene Ontology
Linked Open SGD
Linked Open HGNC
Linked Open PDB
Linked Open Pfam
Linked Open ChEMBL
Linked Open OMIM
Linked Open Affymetrix
Linked Open SISVU
Linked Open LODE
Linked Open AGROVOC
Linked Open GEMET
Linked Open Drug Bank
Linked Open LinkedCT
Linked Open Enipedia
Linked Open Daily Med
Linked Open Disease Ome
Linked Open SIDER
Linked Open STITCH
Linked Open ProDom
Linked Open PRO-SITE
Linked Open Cornetto
Linked Open WordNet (W3C)
Linked Open WordNet (VUA)
Linked Open WordNet (RKB Explorer)
Linked Open YAGO
Linked Open UMBEL
Linked Open Uberblic
Linked Open dbpedia lite
Linked Open Geo Species
Linked Open Fishes of Texas
Linked Open Taxon Concept
Linked Open Geo Linked Data
Linked Open Eurostat
Linked Open Eurostat (Ontology Central)
Linked Open TWC LOGD
Linked Open transport data.gov.uk
Linked Open data.gov.uk intervals
Linked Open legislation data.gov.uk
Linked Open Ordnance Survey
Linked Open UK Post-codes
Linked Open GovWILD
Linked Open LOTUS
Linked Open Statistics data.gov.uk
Linked Open Openly Local
Linked Open OpenEI
Linked Open education.data.gov.uk
Linked Open BBC Programmes
Linked Open BBC Music
Linked Open BBC Wildlife Finder
Linked Open Chronicling America
Linked Open Event Media
Linked Open Linked MDB
Linked Open Portuguese DBpedia
Linked Open Greek DBpedia
Linked Open URI Burner
Linked Open New York Times
Linked Open Telegraphis
Linked Open Rechtspraak.nl
Linked Open Open Calais
Linked Open iServe
Linked Open OS
Linked Open World Factbook
Linked Open Project Gutenberg
Linked Open TCM Gene DIT
Linked Open data-open-ac-uk
Linked Open Revyu
Linked Open RDF Book Mashup
Linked Open DOC
Linked Open BNB
Linked Open data bnf.fr
Linked Open Calames
Linked Open SW Dog Food
Linked Open semantic web.org
Linked Open Sudoc
Linked Open IdRef Sudoc
Linked Open UB Mannheim
Linked Open P20
Linked Open Ulm
Linked Open ndlna
Linked Open PSH
Linked Open GND
Linked Open LCSH
Linked Open LIBRIS
Linked Open VIAF
Linked Open Europeana
Linked Open Swedish Open Cultural Heritage
Linked Open NSZL Catalog
Linked Open Iobid Resources
Linked Open Iobid Organisations
Linked Open STW
Linked Open GESIS
Linked Open RESEX
Linked Open Scholarometer
Linked Open LOCAH
Linked Open DEPLOY
Linked Open IBM
Linked Open IRIT
Linked Open ACM
Linked Open RAE2001
Linked Open Newcastle
Linked Open CiteSeer
Linked Open Eurecom
Linked Open VIVO Cornell
Linked Open VIVO Indiana
Linked Open UN/LOCODE
Linked Open ERA
Linked Open DBLP (RKB Explorer)
Linked Open DBLP (L3S)
Linked Open DBLP (FU Berlin)
Linked Open ECS Southampton EPrints
Linked Open ECS (RKB Explorer)
Linked Open ECS Southampton
Linked Open BibBase
Linked Open OAI
Linked Open Wiki
Linked Open Budapest
Linked Open Pisa
Linked Open Roma
Linked Open NVD
Linked Open Deutsche Biographie
Linked Open Ulm
Linked Open Norwegian MeSH
Linked Open RÅdata nå!
Linked Open LEM
Linked Open t4gm info
Linked Open NDL subjects
Linked Open Open Library (Talis)
Linked Open Open Library
Linked Open NTU Resource Lists
Linked Open St. Andrews Resource Lists
Linked Open Sussex Reading Lists
Linked Open Manchester Reading Lists
Linked Open Plymouth Reading Lists
Linked Open LinkedCT
Linked Open LinkedL CCN
Linked Open Thesaurus W
Linked Open MARC Codes List
Linked Open theses.fr
Linked Open SSW Thesaurus
Linked Open Source Code Ecosystem Linked Data
Linked Open Ontos News Portal
Linked Open Linked Crunchbase
Linked Open RDF ohloh
Linked Open gnoss
Linked Open Didactalia
Linked Open my Experiment
Linked Open flickr wrapper
Linked Open Good-win Family
Linked Open Pokédex
Linked Open Poképédia
Linked Open Classical (DB Tune)
Linked Open Jamendo (DBtune)
Linked Open Music Brainz (zitgist)
Linked Open Music Brainz (DBTune)
Linked Open Music Brainz (Data Incubator)
Linked Open Last.FM (rdfize)
Linked Open Last.FM artists (DBTune)
Linked Open Discogs (Data Incubator)
Linked Open Surge Radio
Linked Open FanHubz
Linked Open patents data.go v.uk
Linked Open research data.gov.uk
Linked Open business data.gov.uk
Linked Open Crime (En-AKTing)
Linked Open Mortality (En-AKTing)
Linked Open Population (En-AKTing)
Linked Open NHS (En-AKTing)
Linked Open CO2 Emission (En-AKTing)
Linked Open Energy (En-AKTing)
Linked Open Ren. Energy Generators
Linked Open EEA
Linked Open reegle
Linked Open Ox Points
Linked Open Crime Reports UK
Linked Open Hellenic PD
Linked Open Hellenic FBD
Linked Open EUTC Productions
Linked Open DB Tropes
Linked Open Magnatune
Linked Open John Peel (DBTune)
Linked Open GTAA
Linked Open Moseley Folk
Linked Open Audio Scrobbler (DBTune)
Linked Open LOV
Linked Open User Feedback
Linked Open Slideshare 2RDF
Linked Open tags2con delicious
Linked Open Bricklink
Linked Open yovisto
Linked Open Semantic Tweet
Linked Open Lotico
Linked Open Klapp-stuhl-club
Linked Open Data Gov.ie
Linked Open Traffic Scotland
Linked Open Scotland Pupils & Exams
Linked Open ISTAT Immigration
Linked Open Brazilian Politicians
Linked Open EU Institutions
Linked Open Open Election Data Project
Linked Open Lichfield Spending
Linked Open ESD standards
Linked Open reference data.gov.uk
Linked Open London Gazette
Linked Open CORDIS (RKB Explorer)
Linked Open CORDIS (FUB)
Linked Open GovTrack
Linked Open EURES
Linked Open Linked EDGAR (Ontology Central)
Linked Open FTS
Linked Open Scotland Geography
Linked Open Finnish Municipalities
Linked Open Piedmont Accomodations
Linked Open Italian public schools
Linked Open El Viajero Tourism
Linked Open Turismo de Zaragoza
Linked Open Semantic XBRL
Linked Open US SEC (rdfabout)
Linked Open GeoWordNet
Linked Open SMC Journals
Linked Open Ocean Drilling Codices
Linked Open Janus AMP
Linked Open Yahoo! Geo Planet
Linked Open AEMET
Linked Open Climbing
Linked Open Twarql
Linked Open EUNIS
Linked Open US Census (rdfabout)
Linked Open riese
Linked Open Eurostat (Ontology Central)
Linked Open Open Cyc
Linked Open totl.net
Linked Open Laxvo
Linked Open Metoffice Weather Forecasts
Linked Open EARTh
Linked Open Alpine Ski Austria
Linked Open Weather Stations
Linked Open Airports
Linked Open Product DB
Linked Open Product Types Ontology
Linked Open Italian Museums
Linked Open Google Art wrapper
Linked Open Smart Link
Linked Open Open Data Thesaurus
Linked Open Sears
Linked Open medu-cator
Linked Open National Radioactivity JP
LINKED OPEN DATA
On the web
Machine-readable data
Non-proprietary format
RDF standards
Linked RDF
IS YOUR DATA 5 ★?
As of September 2011

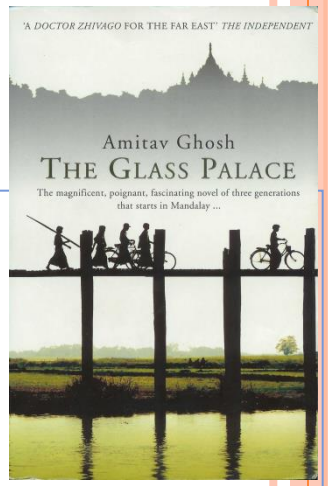
EXAMPLE: LINKED DATA SCENARIO



This example will show how data from two *bookstores* can be exposed and published in a structured way, how they can be **integrated**, how they can be **queried**, how they can be **enriched** with other knowledge (e.g. context or domain knowledge) from other sources (e.g. wikipedia).

See acknowledgements for the source of this example

Linked Data Scenario
A simplified bookstore data (dataset "A")



Suppose the data of the bookstore A are stored in a relational database. The relational tuples related to the book of our running example could be

Books

ISBN	Author	Title	Publisher	Year
0006511409X	id_xyz	The Glass Palace	id_qpr	2000

Authors

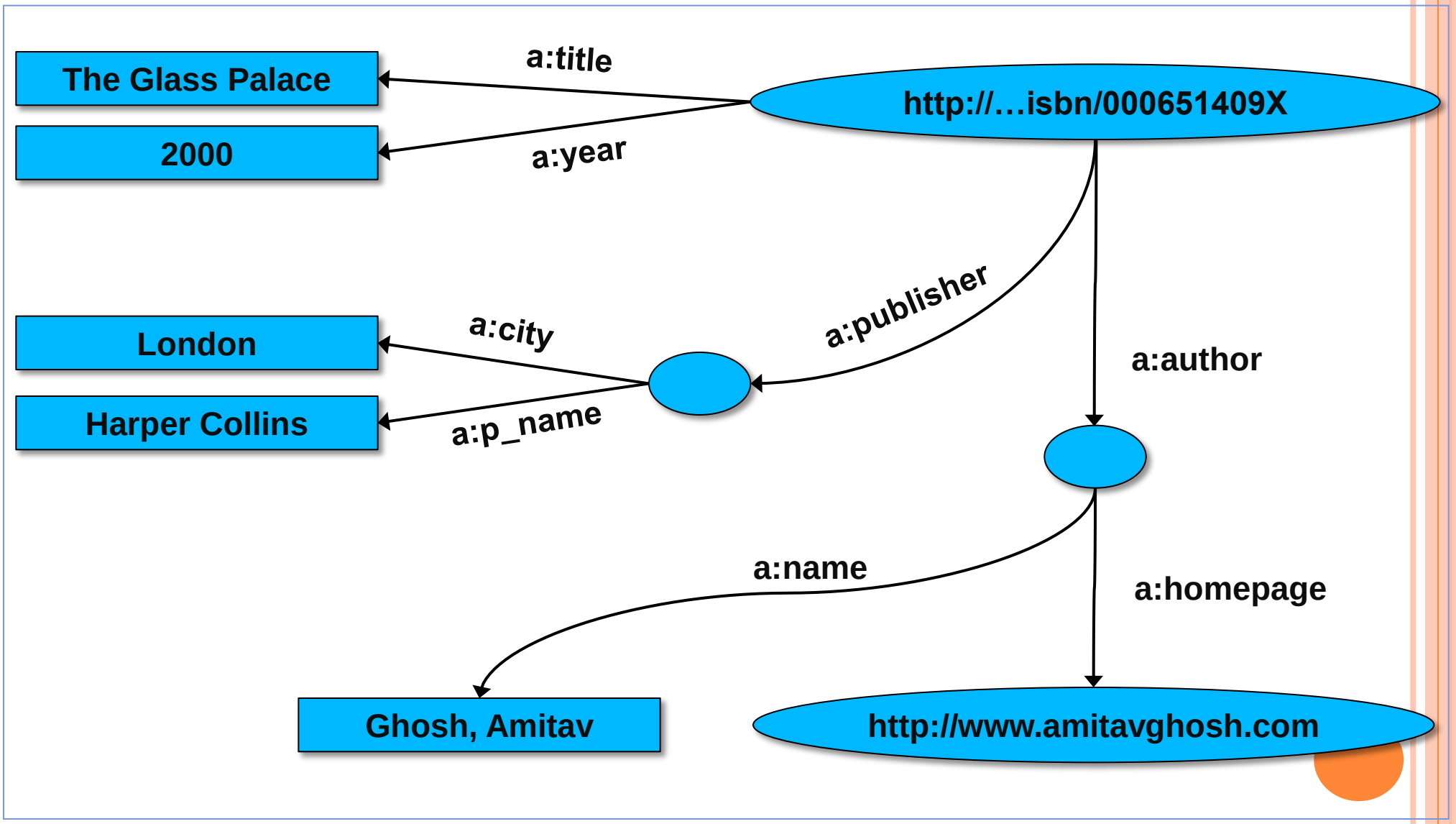
ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Publishers

ID	Publisher's name	City
id_qpr	Harper Collins	London

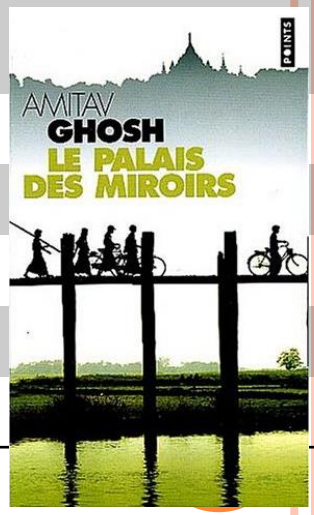
LINKED DATA SCENARIO

1ST: EXPORT YOUR DATA AS A *GRAPH*

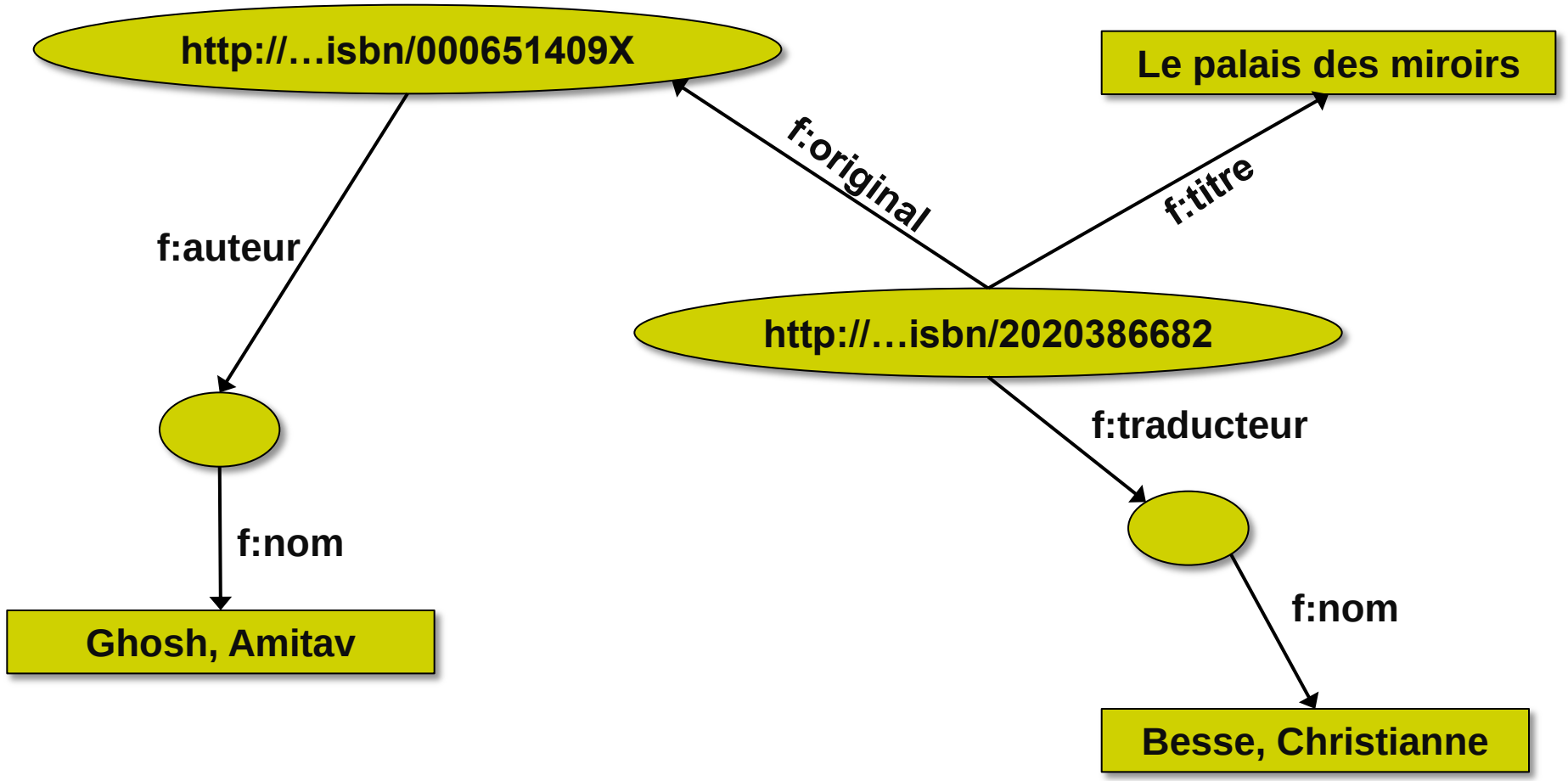


Linked Data Scenario
Another bookstore data (dataset "F" from France)

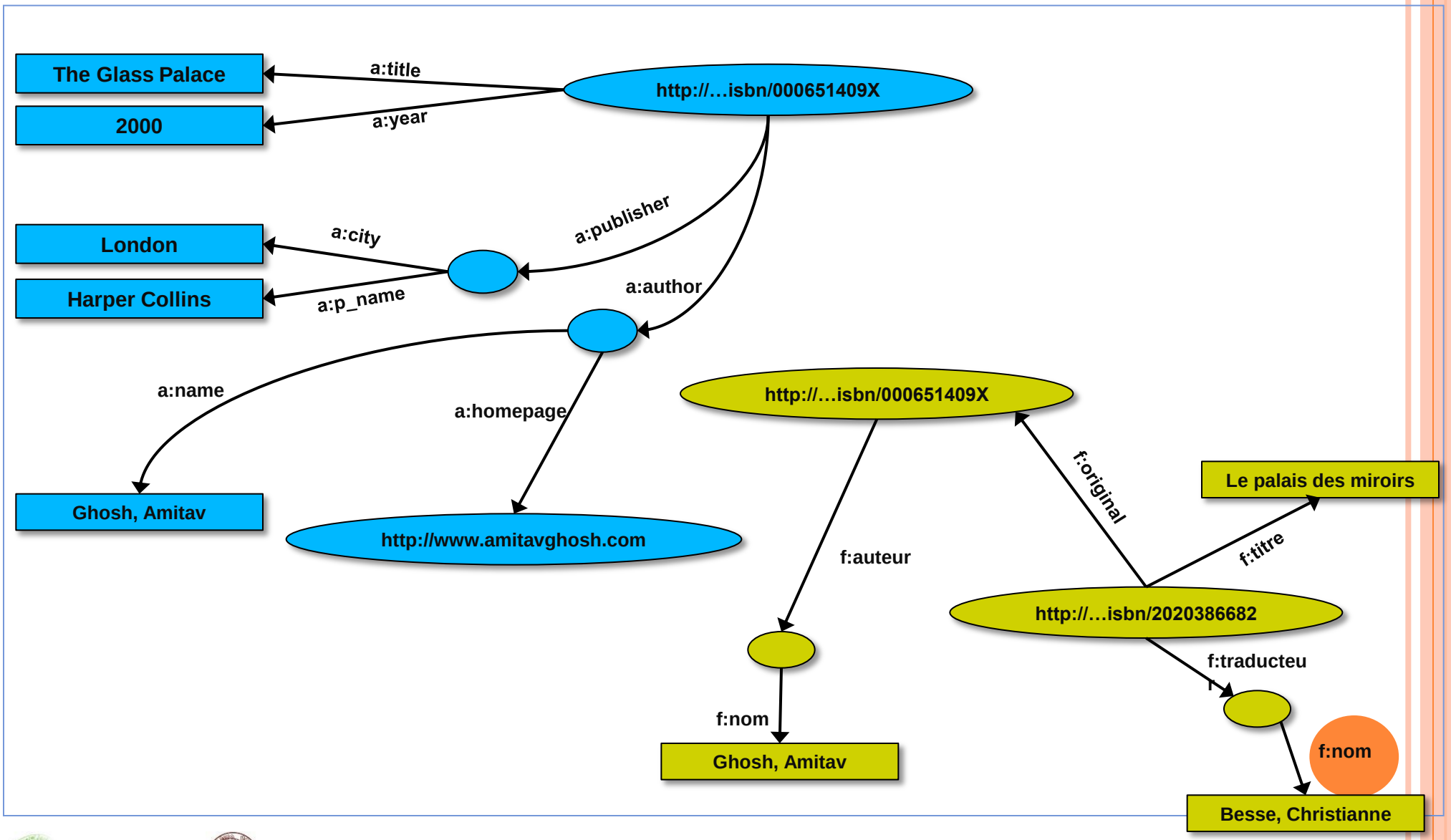
	A	B	C	D
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN 2020286682	Le Palais des Miroirs	$A12$	ISBN 0-00-6511409-X
3				
4				
5				
6	**ID**	**Auteur**		
7	ISBN 0-00-6511409-X	$A11$		
8				
9				
10	**Nom**			
11	Ghosh, Amitav			
12	Besse, Christianne			



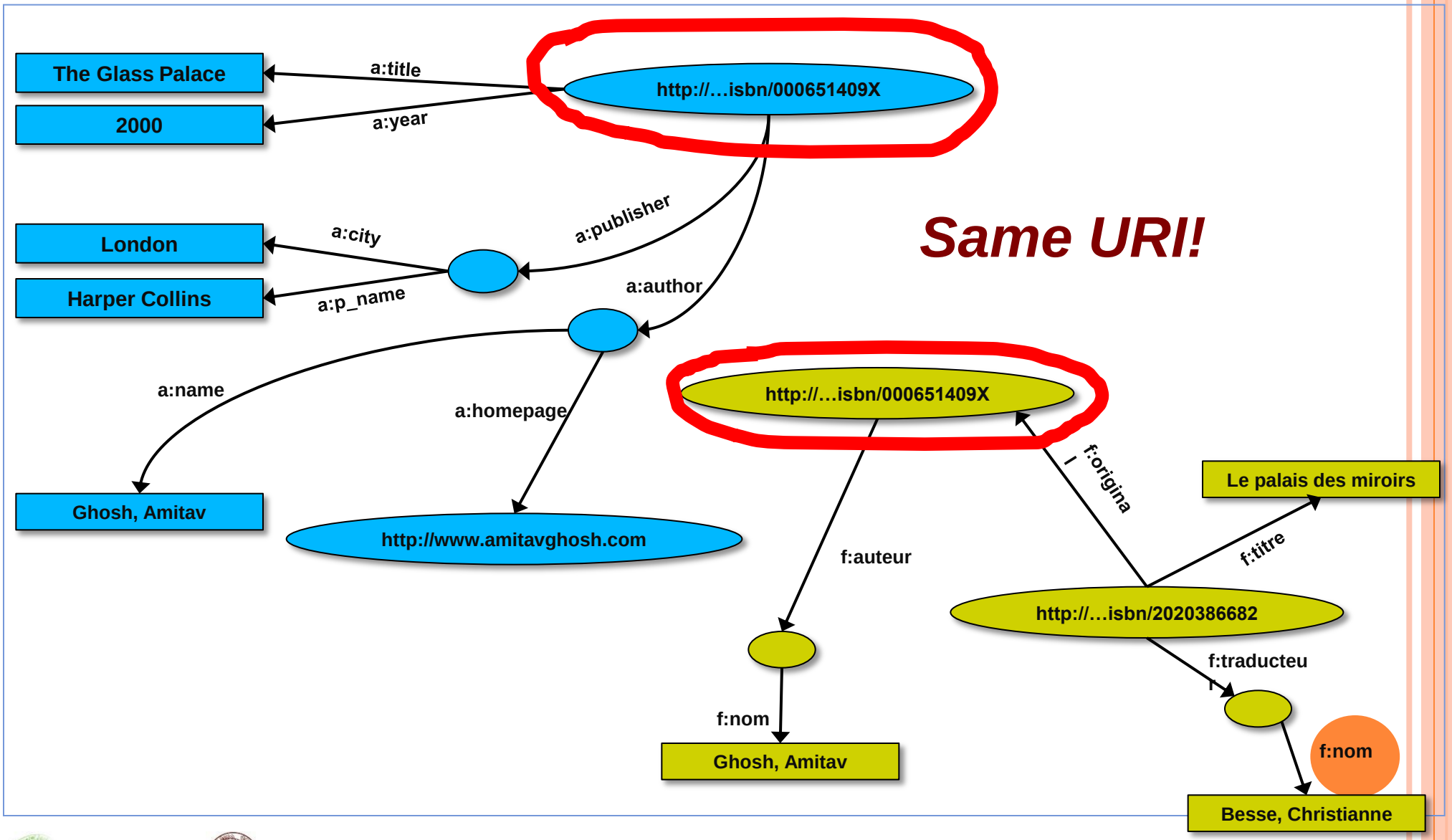
Linked Data Scenario
2ND: EXPORT YOUR SECOND SET OF DATA



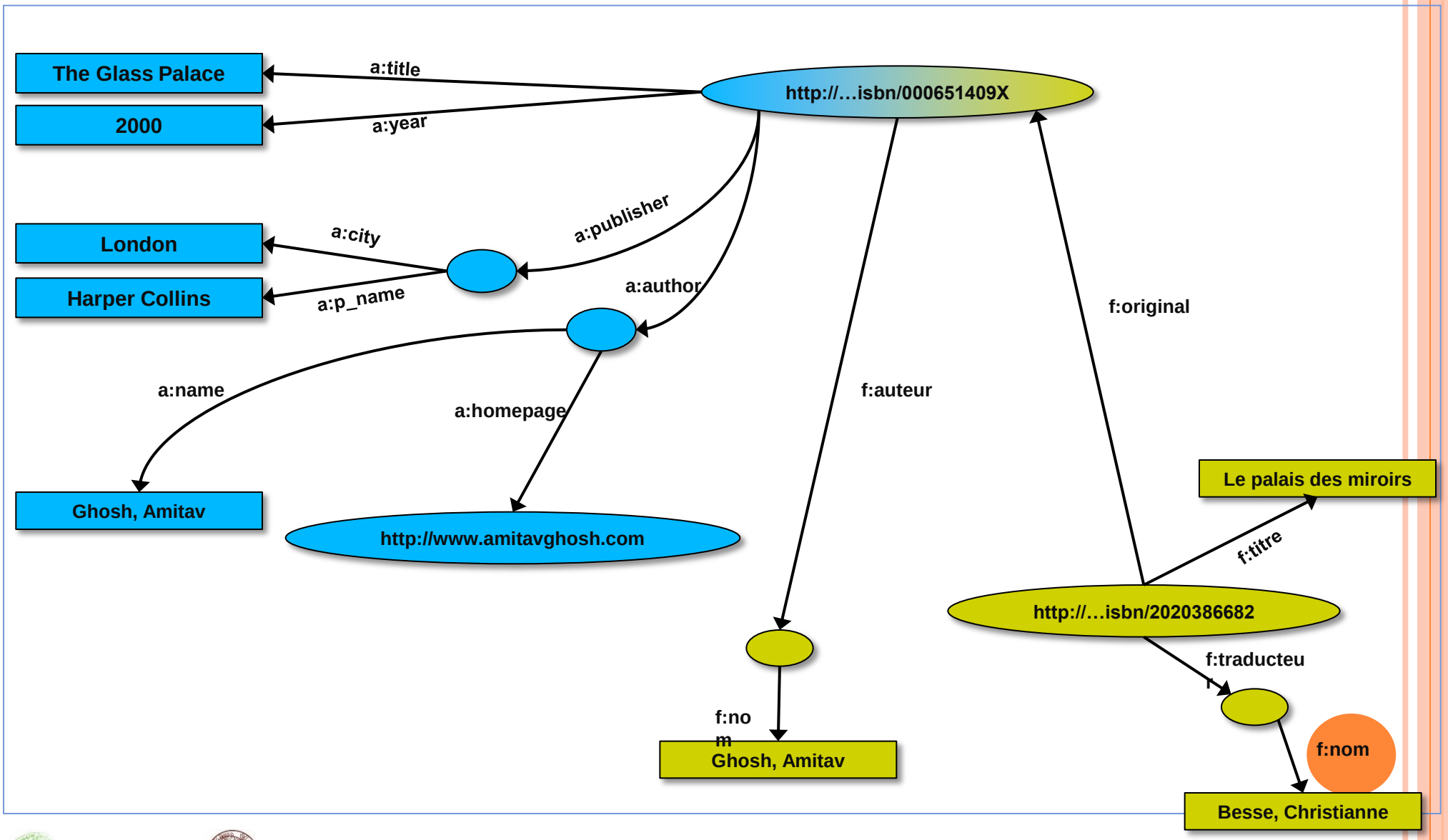
LINKED DATA SCENARIO
3RD: START MERGING YOUR DATA



Linked Data Scenario
3RD: start merging your data (cont)

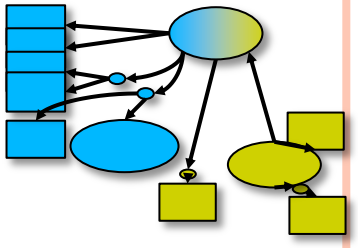


Linked Data Scenario
3RD: start merging your data



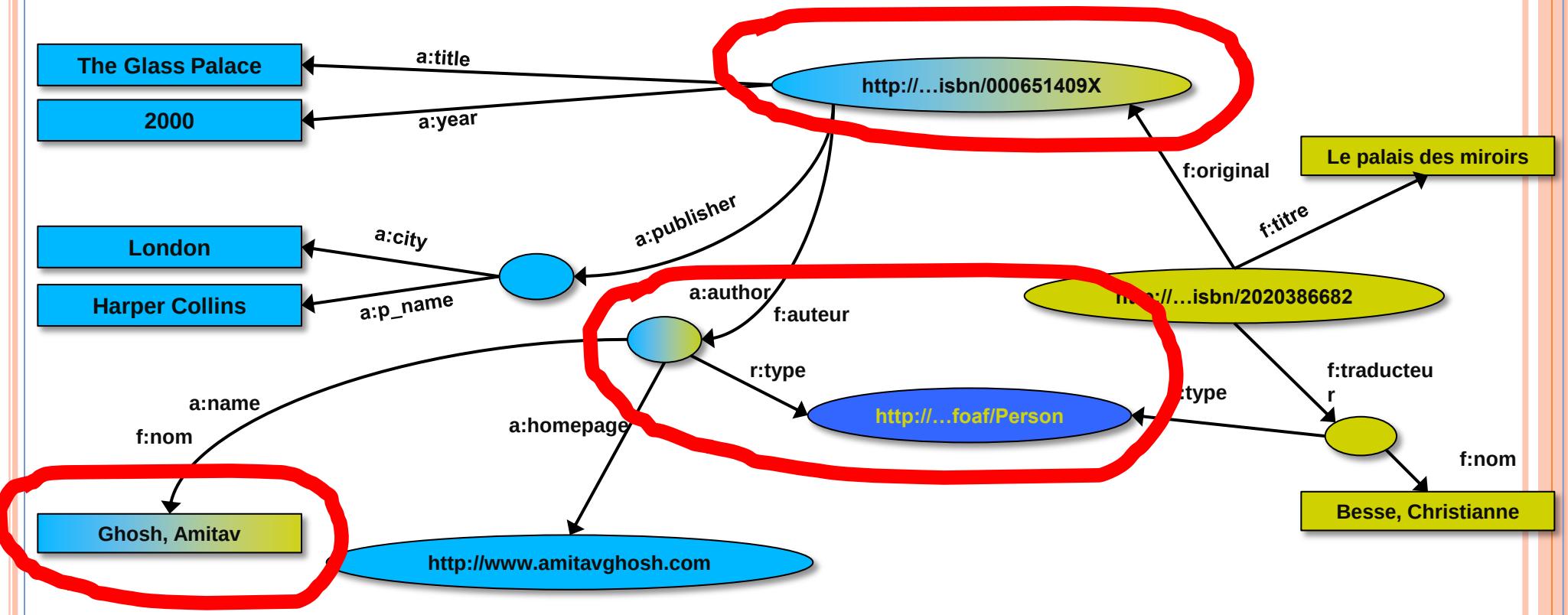
LINKED DATA SCENARIO
START MAKING QUERIES…

- User of data "F" can now ask queries like:
  - ***"give me the title of the original"***
- This information is not in the dataset "F", but it can be retrieved by merging with dataset "A"



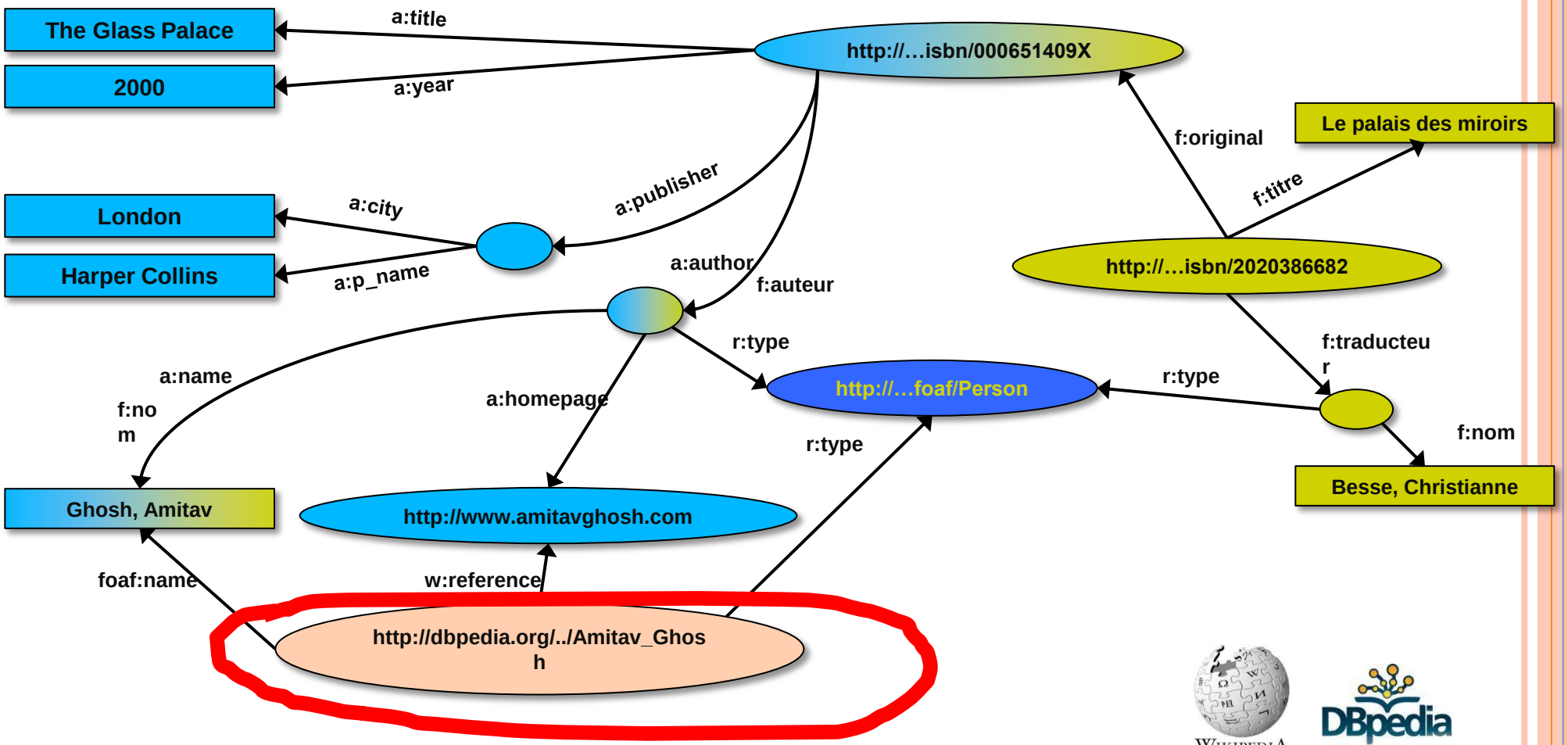
- We "feel" that `a:author` and `f:auteur` should be the same. But an automatic merge does not know that! We can add some extra information to the merged data:
  - `a:author same as f:auteur`
  - both identify a "Person" a term that a community may have already defined (a "Person" is uniquely identified by his/her name and, say, homepage (it can be used as a "category" for certain type of resources)

Linked Data Scenario
3rd revisited: use the extra knowledge



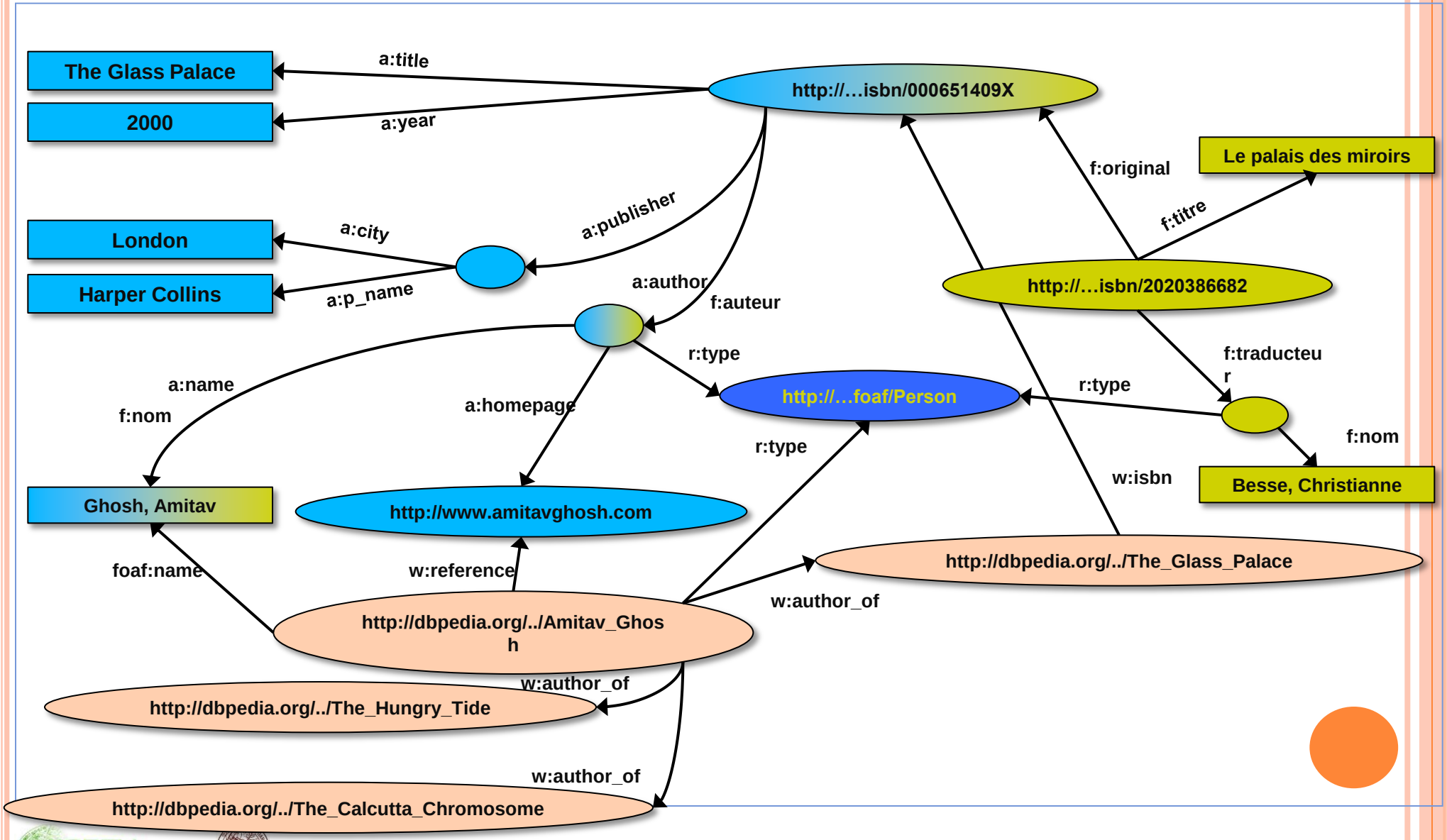
- User of dataset "F" can now query: "***give me the home page of the original's 'auteur'***". The information is not in datasets "F" or "A", but was made available by **merging** datasets "A" and datasets "F" and adding three simple extra statements as an extra "glue"

Linked Data Scenario
Merge with Wikipedia data

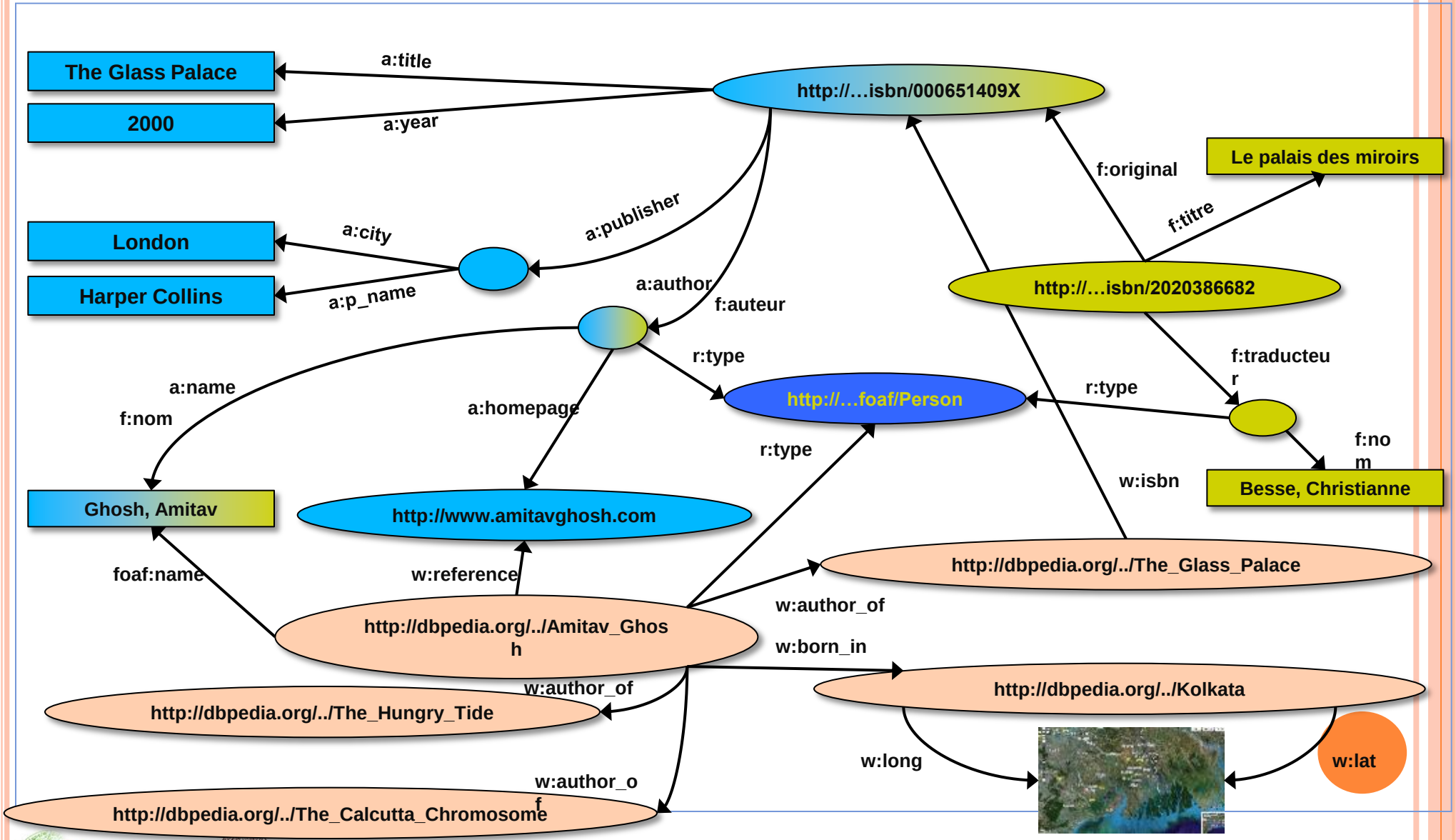


- Using, e.g., the "Person", the dataset can be combined with other sources. For example, data in Wikipedia can be extracted using dedicated tools

Linked Data Scenario
Merge with Wikipedia data



Linked Data Scenario
Merge with Wikipedia data

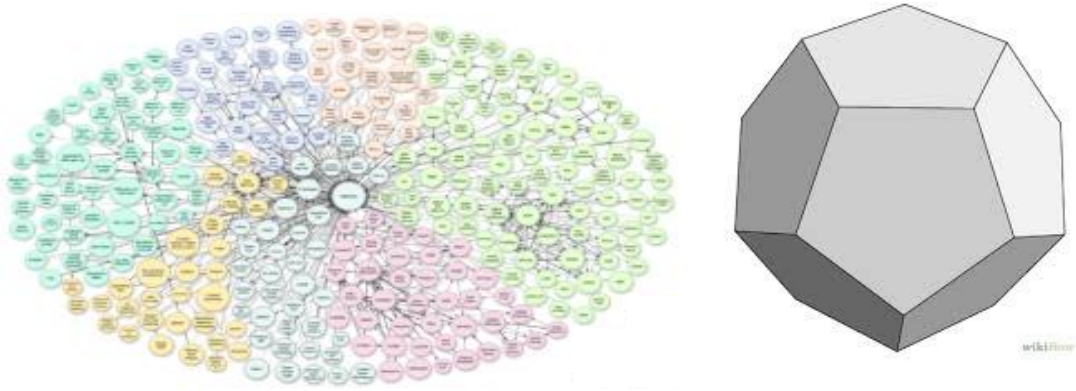


Queries and The Difficulty of Query Formulation

- The availability of Semantic Data allows formulating complex queries. E.g.:
  - ***scientists who have worked in an institute in Germany and are known for their work in fuzzy set theory***
- however to formulate such queries one has to know not only SPARQL, but also the vocabulary used. E.g. (sketch):

```
SELECT ?x
WHERE {
    ?x type scientist .
    ?x ns:knownFor “fuzzySetTheory” .
    ?x ns:hasWorkedIn ?y .
    ?y type institution .
    ?y ns:locatedAt “Germany” .
}
```

QUERYING THE SEMANTIC WEB / LOD



- Faceted exploration can allow users to satisfy their information needs without having to be aware of the **employed terminology**, **contents**, or **query language** of the sources

- An example from an application offering faceted exploration to DBpedia follows

German scientists known for their work in the field of evolution

Guided exploration - An example

Transition markers

Surveying Faceted Exploration Approaches

We use the term *BA* to refer to a *browsing approach*. We can survey the works (*BA*s) that have been applied or proposed according to various aspects:

1. ***Characteristics of the underlying information space***. The structuring of the underlying information base is an important aspect since each case requires tackling different difficulties.
2. ***Configuration***. Some approaches can be applied without requiring any form of configuration or application design (regarding the browseable information space), while some others require configuration steps, e.g. specify the contents and structuring of the browsable part through the view-based approach over a DB or an RDF repository. Since the browsable part of the information source is defined by a query, its structure may be different from that of the original source.
3. ***State Space***. In general we can view the interaction as a state space consisting of *states* and *transitions*, therefore we can characterize, or comparatively evaluate, two *BA*s by comparing their state spaces, e.g. by identifying properties which are satisfied by their state space.

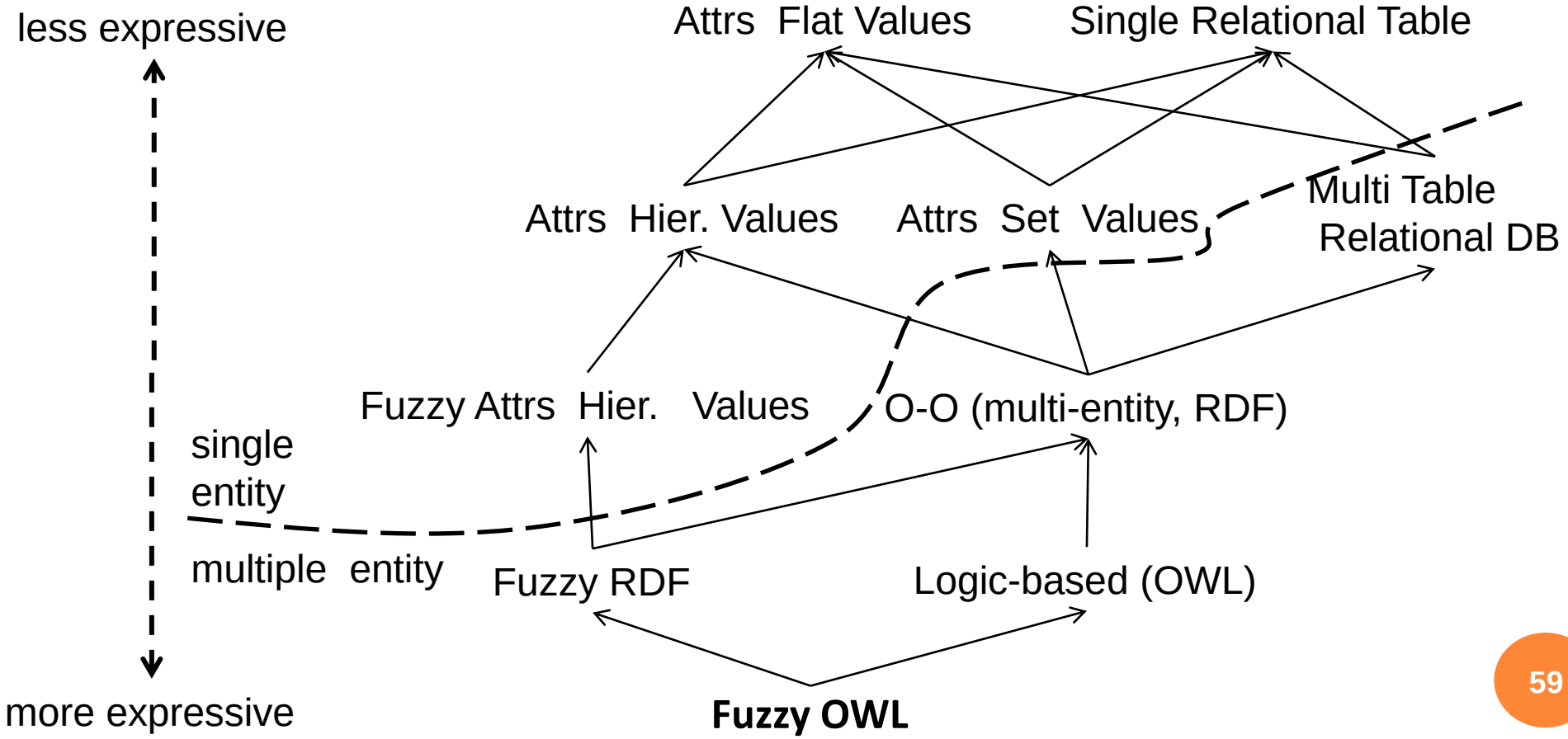
Surveying Faceted Exploration Approaches

1. *THE UNDERLYING INFORMATION SPACE*

- Some browsing approaches are applicable to simple structures, while others to complex information structures (e.g. OWL-based KBs). There are several options, some of them follow:
  - **attribute-value pairs** with **flat** values (e.g. name=Yannis),
  - **attribute-value pairs** with **hierarchically organized values** (e.g. location=Crete),
  - **set-valued attributes** (either flat or hierarchical) (e.g. accessories={ABS, ESP}),
  - **multi-entity** (or object-oriented) (e.g. RDF, linked open data), and relational databases.
  - Furthermore, we could have **fuzziness** and we can consider this as an independent aspect (e.g. there are fuzzy extensions of the RDF model).
- Therefore one important aspect is how the underlying information is structured.

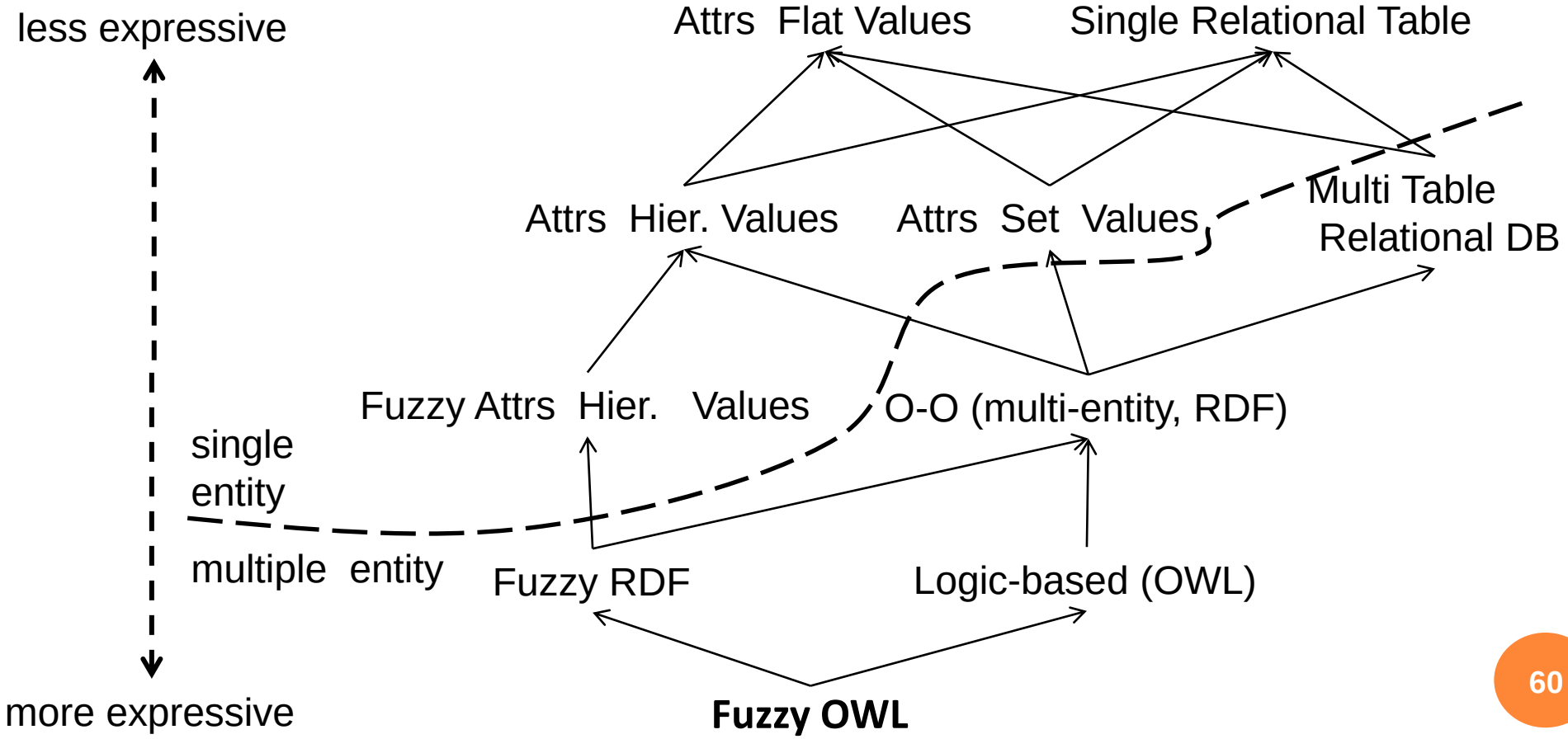
CONT.

Here we can see the above categories **organized hierarchically** where an option X is a (direct or indirect) child of an option Y if whatever information can be expressed in Y can also be expressed in X.



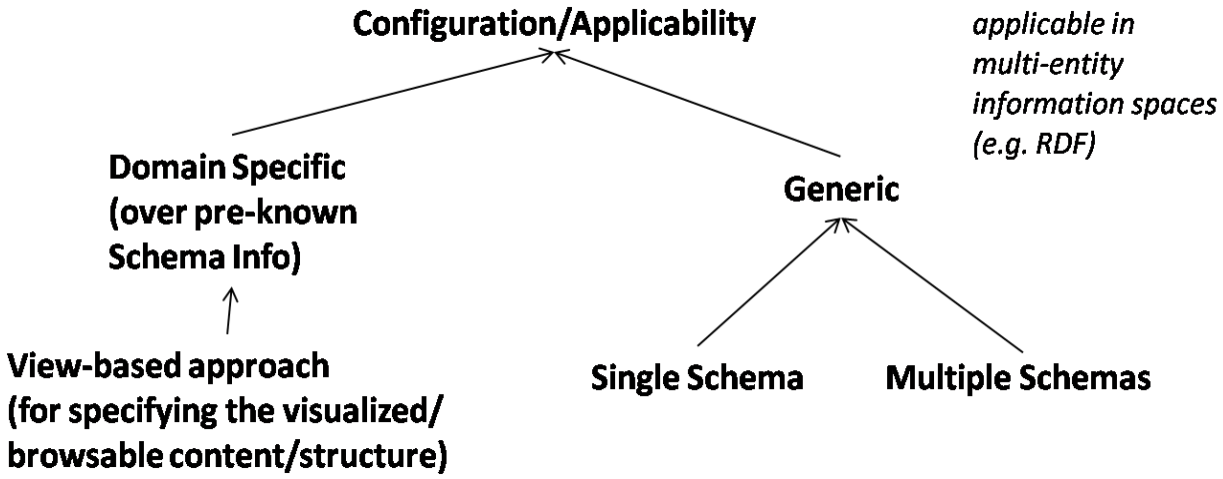
CONT.

The **value of this diagram** is that it depicts the fact that if a browsing approach is appropriate for an option X, then certainly it is appropriate for all options which are parents of X.



Surveying Faceted Exploration Approaches
2. *Configuration*

- *View-based* approach
  - The structure of the browsable part has to be explicitly specified through configuration steps (*e.g.* by using a query language or logic rules)
- *Generic* approach
  - No configuration requirements w.r.t the underlying information space

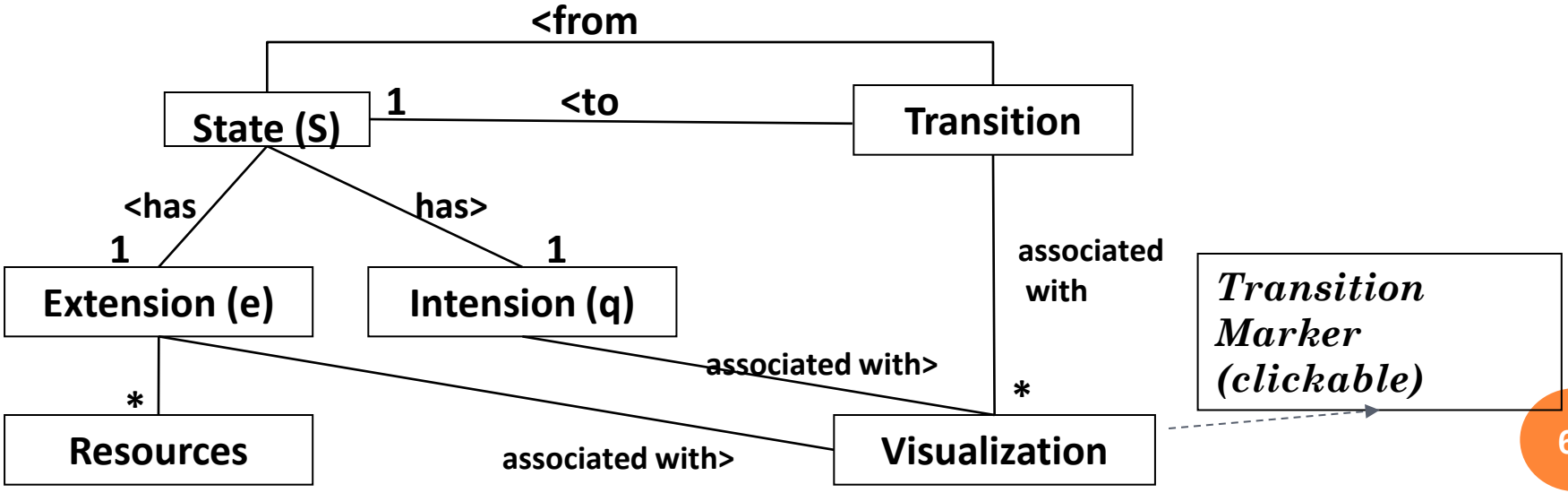


Surveying Faceted Exploration Approaches
3. State Space (States, Transitions and Transition Markers)

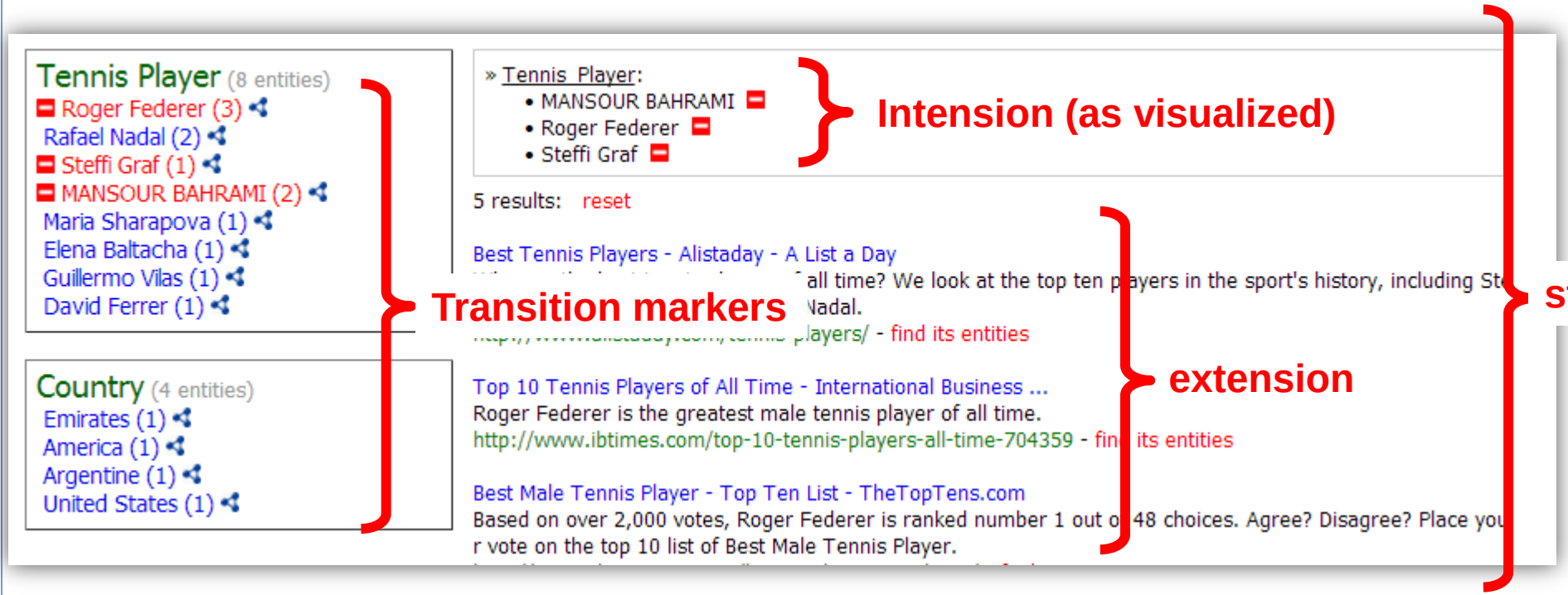
- **State Space**. In general we can view the interaction as a state space consisting of **states** and **transitions**, therefore we can characterize, or comparatively evaluate, two BAs by comparing their state spaces, e.g. by identifying properties which are satisfied by their state space.

CONT. STATES

- A **state** has an **extension** (a set of items displayed, e.g. WS results), an **intension** (i.e. condition or query; satisfied by the extension), probably a name, and a number of **transitions** each leading to a different state.
- In addition each state has a visualization format for its (a) extension, (b) intention, as well as (c) its transitions (e.g. a tree-control, a list, a table). In any case, each transition has a clickable **transition marker** signifying the existence of the transition. Usually these markers are enriched with information regarding the target state.

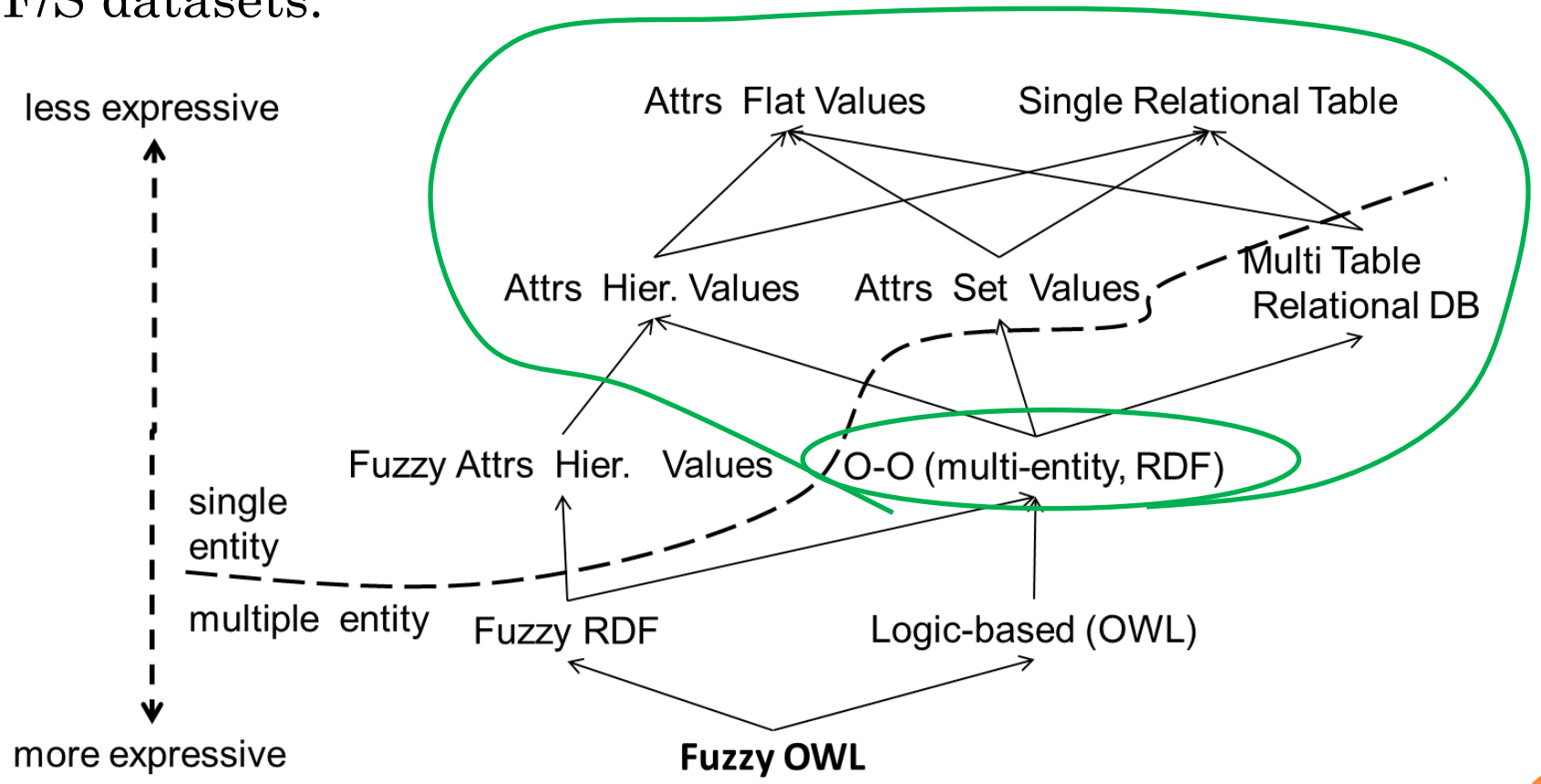


CONT. STATES


Tennis Player (8 entities)
Roger Federer (3)
Rafael Nadal (2)
Steffi Graf (1)
MANSOUR BAHRAMI (2)
Maria Sharapova (1)
Elena Baltacha (1)
Guillermo Vilas (1)
David Ferrer (1)
Country (4 entities)
Emirates (1)
America (1)
Argentine (1)
United States (1)
Transition markers
» Tennis Player:
• MANSOUR BAHRAMI
• Roger Federer
• Steffi Graf
Intension (as visualized)
5 results: reset
Best Tennis Players - Alistaday - A List a Day
all time? We look at the top ten players in the sport's history, including St
Nadal.
layers/ - find its entities
Top 10 Tennis Players of All Time - International Business ...
Roger Federer is the greatest male tennis player of all time.
http://www.ibtimes.com/top-10-tennis-players-all-time-704359 - fin its entities
Best Male Tennis Player - Top Ten List - TheTopTens.com
Based on over 2,000 votes, Roger Federer is ranked number 1 out o 48 choices. Agree? Disagree? Place you
r vote on the top 10 list of Best Male Tennis Player.
extension
state

A "Generic" Interaction Model for Exploring RDF/S Datasets

- Now we will see an interaction model for faceted exploration over RDF/S datasets.



A Generic Interaction Model for Exploring RDF/S Datasets

- Objective: Through simple clicks the user can reach states whose extension corresponds to the answer of complex queries.
- Example:

Japanese cars for sale which are driven by persons who work at FORTH and know a person who knows Bob

- Source: It is a simplified version of the model for browsing Fuzzy RDF that is described in
  - Nikos Manolis, Yannis Tzitzikas: Interactive Exploration of Fuzzy RDF Knowledge Bases. ESWC (1) 2011: 1-16

Supported Kinds of Transitions

- Supported transitions
  - Class-based
  - Property-based
  - Property-path based
  - Entity Type Switch

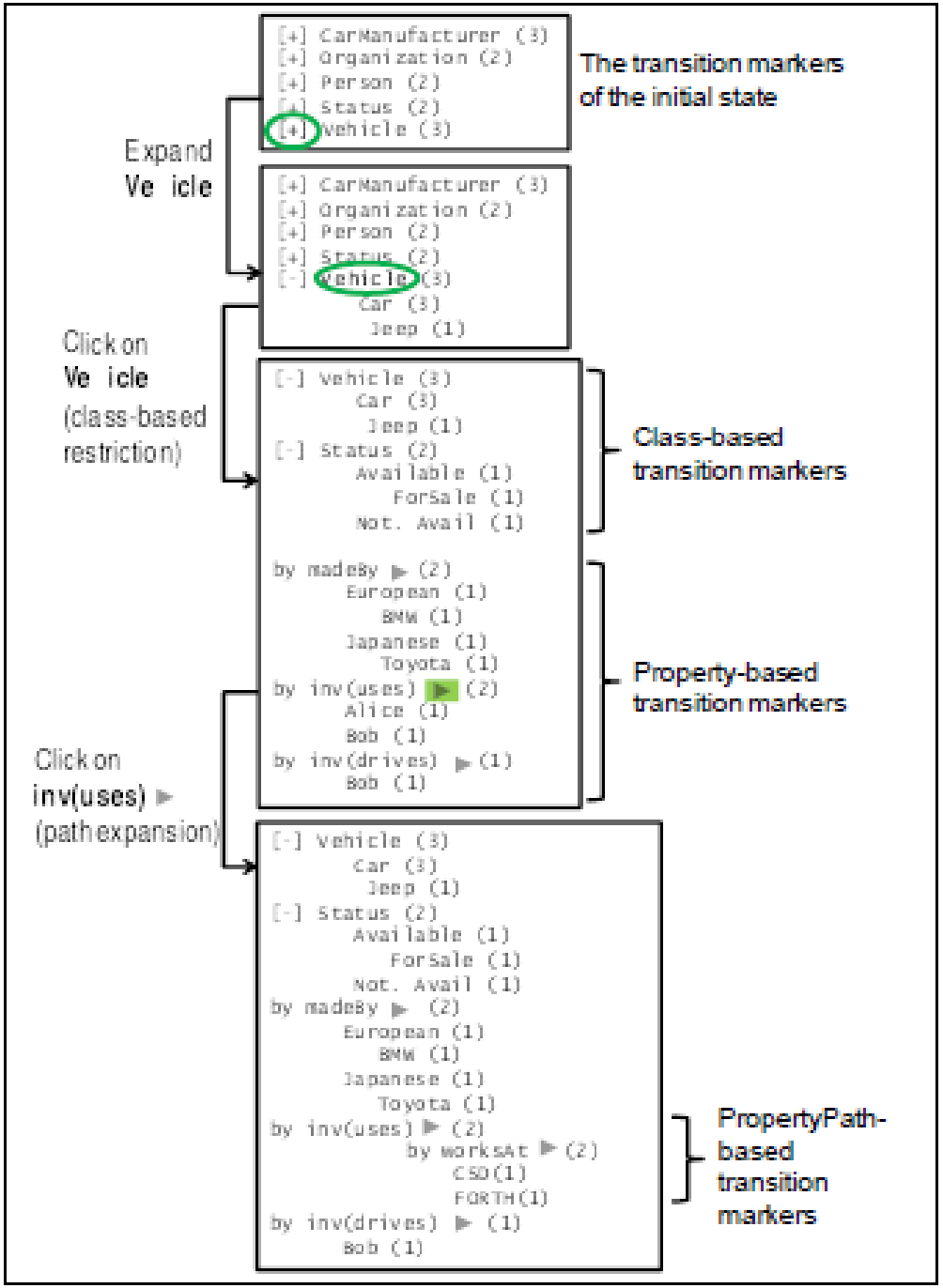
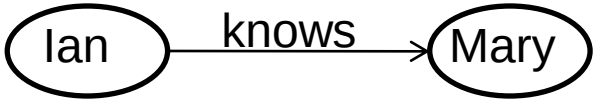


Fig. 10. Sketch of the GUI part for transition markers

NOTATIONS (FOR RDF/S)

- A RDF/S KB is defined by a set of **RDF triples** of the form:

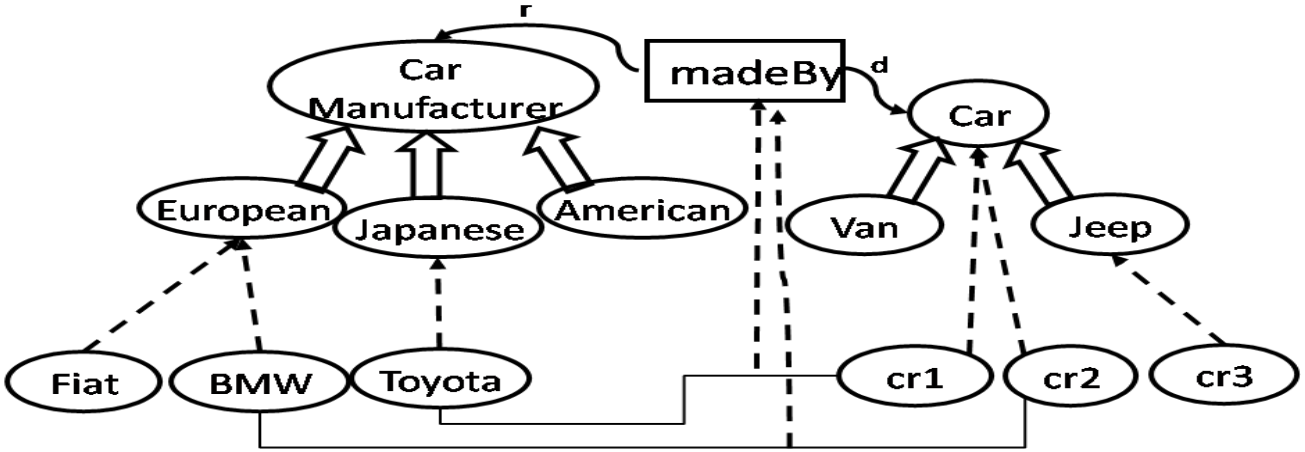
(subject, predicate, object)



- The ***closure*** *C(K)* of a KB K contains all the triples explicitly asserted or inferred (based on the RDF/S semantics) from a KB

- The ***Schema*** of a RDF/S KB K is a 6-tuple Γ=<C, Pr, *domain, range,* $\leq^*_{cl}$, $\leq^*_{pr}$ >
  - *C: Classes, Pr: Propertis, domain and range of properties, subclassOf (among C), subPropertyOf (among Pr)*
- ***Instance notations*** for a KB:
  - *Instances of a class c* ϵ *C:* **inst(c)** = { **o** | (o, type, c) $\in$ *C(K)* }
    - *Class instance triples*: **(o type c)**
  - *Instance of a property* p ϵ *Pr:* **inst(p)** = { **(o, p, o')** | (o, p, o') $\in$ *C(K)* }
    - *Property instance triples*: **(o p o')**

CONT.



- **inst(Jeep)={cr3}**
- **inst(Car)={cr1, cr2, cr3}**
- **Inst((Car,madeBy,CarManufacturer)) = {**
  (cr1, madeBy,Toyota)
  (cr2, madeBy,BMW)
  }

AUXILIARY DEFINITIONS: RESTRICTIONS AND JOINS

Assuming an initial state we need to tackle the following:

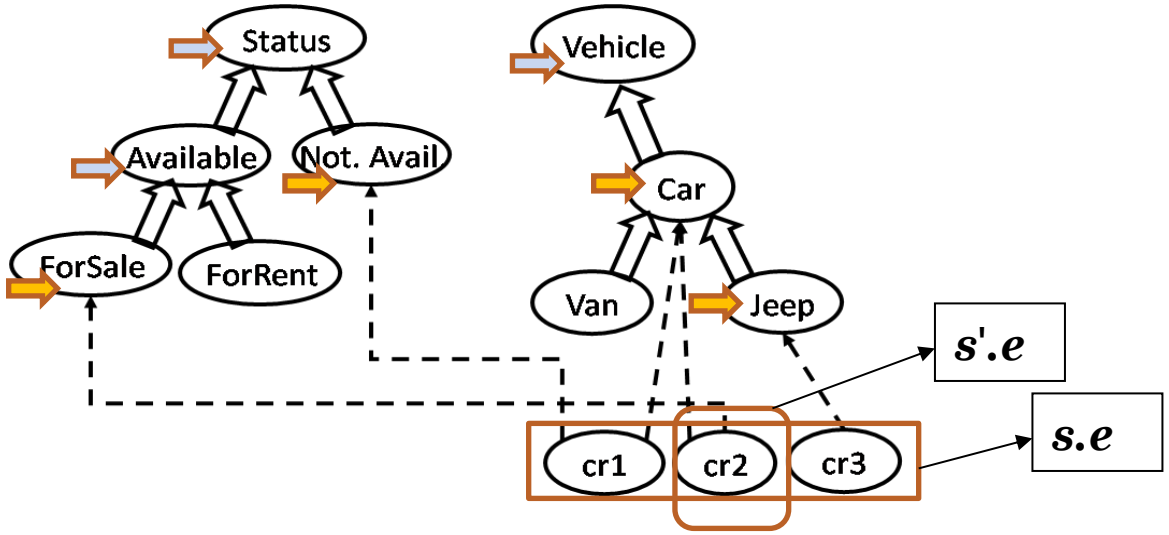
- How the **transitions markers** (*tms*) available to that state are computed
- How the **new state's extension** is computed after selecting a *tm*

- To define the above we need to define the notion *of* ***restriction*** and ***join***

p ϵ PR ∪ PR $^{-1}$, **E**: set of resources, **vset** : a set of resources or literals

- ***Restrictions*** : Given a set E
  1. *Restrict* (E, p : v) = { e ∈ *E* | (*e, p, v*) ∈ *inst*(*p*)}
  2. *Restrict* (E, p : vset) = { e ∈ *E* | *v'* ∈ vset and (*e, p, v'*) ∈ *inst*(*p*)}
  3. *Restrict* (E, c) = { e ∈ E | e ∈ *inst*(*c*)}
- ***Joining values***: Compute values which are linked with the elements of E
  - *Joins* (E, p) = { v | ∃ e ∈ *E and (e, p, v)* ϵ *inst*(*p*)}

CLASS-BASED TRANSITIONS



Candidate tms: $TM_{cl}(s.e) = \{ c \in C \mid Restrict\ (s.e, c) \neq \varnothing \}$

Clicking on a $c \in TM_{cl}(s.e)$ then s'.e = *Restrict (s.e, c)*

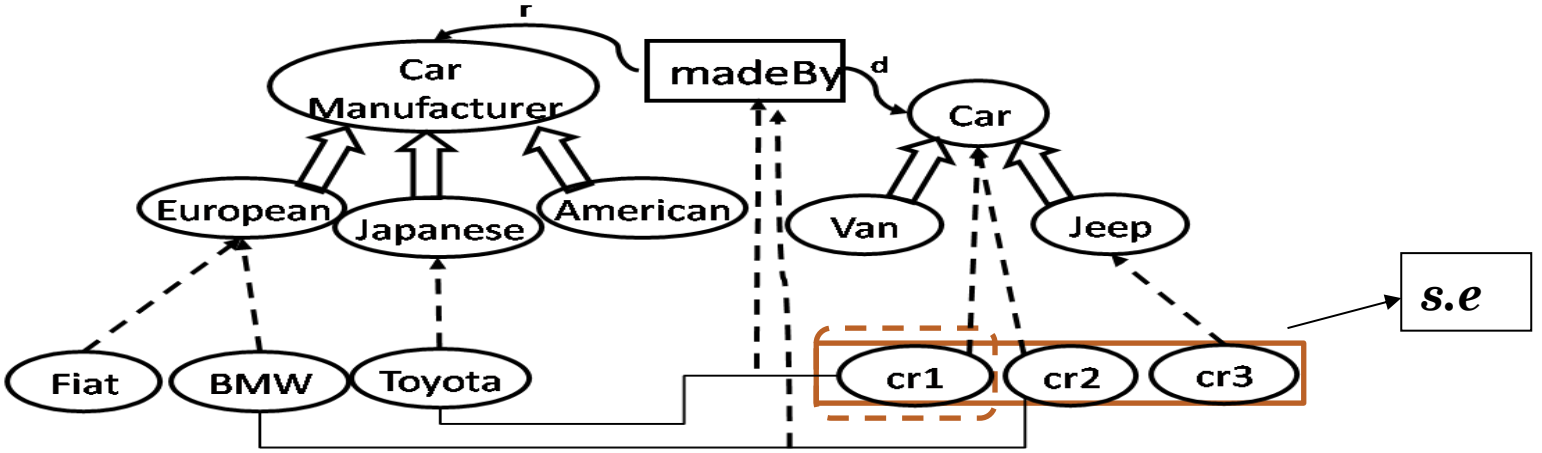
[-]Vehicle(3)
  Car(3)
    Jeep(1)
[-] Status(2)
  Available(1)
    ForSale(1)
  Not Available(1)

→

[-] Vehicle(1)
  Car(1)
[-] Status(1)
  Available(1)
    ForSale(1)

PROPERTY-BASED TRANSITIONS

- The goal: Restrict the current state's extension according to a property value **v** (resource or literal). E.g. All cars made by Toyota



- Candidate properties: Props(s) = { p∈ Pr ∪ PR^{-1} | Joins(s.e, p) ≠∅ }

Candidate tms for a p∈ Props(s) : Joins (s.e, p) = { v | e ∈ s.e and (e, p, v) ϵ inst(p)}

by madeBy(2)
BMW(1)
Toyota(1)

OR

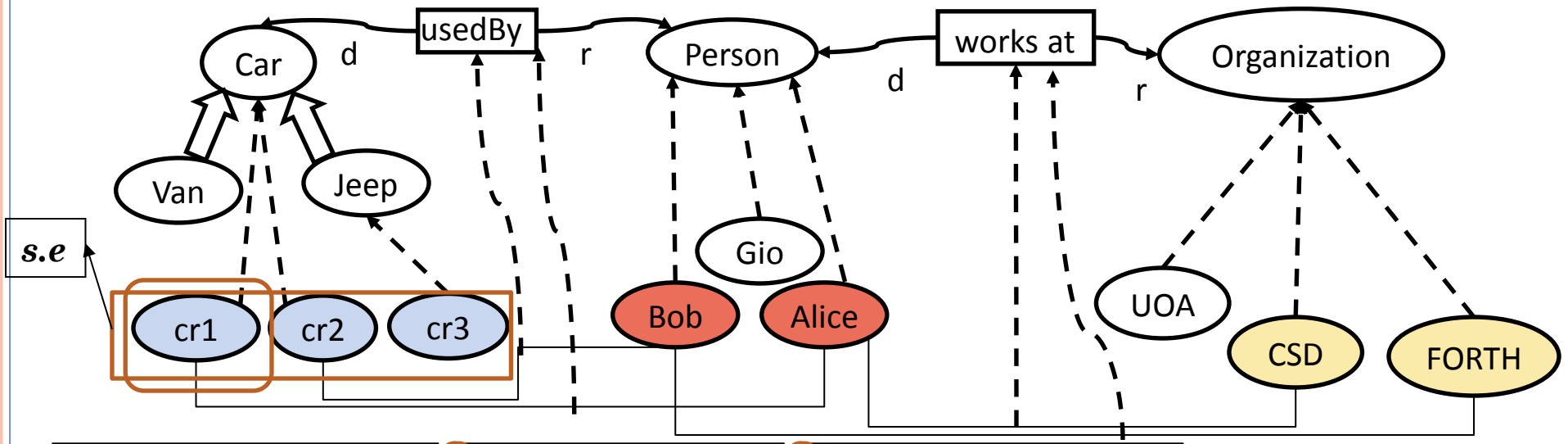
by madeBy(2)
European
BMW(1)
Japanese (1)
Toyota(1)

TM_{cl} (*Joins* (s.e, p))

Property Path-Based Transitions

- Let $p_1, \ldots, p_k$ be a sequence of properties

  E.g. All cars used by persons working at CSD



s.e = M_0 = {cr1,cr2, cr3} → M_1 = {Bob, Alice} → M_2 = {CSD, FORTH}

$M_i = Joins(M_{i-1}, p_i)$

s'.e = M'_0 = {cr1} → M'_1 = {Alice} → M'_2 = {CSD}

$M'_i = Restrict(M_i, p_{i+1} : M'_{i+1})$

by usedBy(2)
Alice (1)
Bob (1)

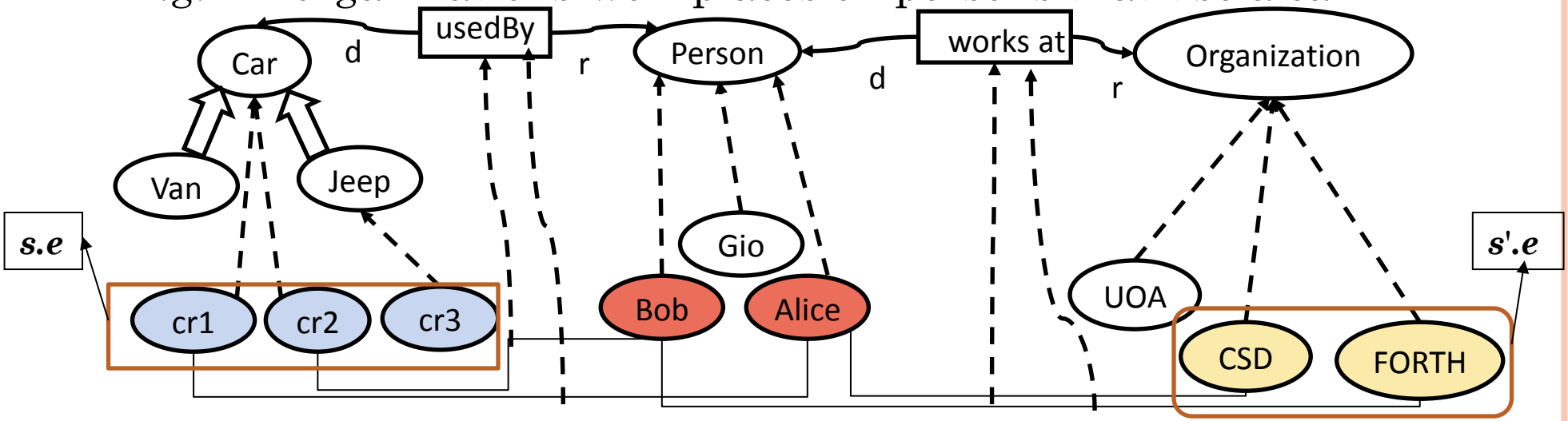
Path expansion

by usedBy(2)
by worksAt(2)
CSD(1)
FORTH(1)

ENTITY TYPE SWITCH TRANSITIONS

Allow users to move to a state whose extension is the current set of tms

E.g. All organizations-workplaces of persons that use a car



s.e = M_0 = {cr1,cr2, cr3} → M_1 = {Bob, Alice} → M_2 = {CSD, FORTH}

s'.e = M_2 = {CSD, FORTH}

by inv(uses)(2)
all
Alice (1)
Bob (1)

Path expansion

by inv(uses)(2)
by worksAt(2)
all
CSD(1)
FORTH(1)

APPLICABILITY

- To apply it, one has to implement Restrict and Join using the technology used for the underlying dataset
  - (e.g. over SPARQL, over SQL, over WhatEverQL, …)
- For more see the related publication
  - It also captures the fuzzy aspect

Synopsis of the first Part

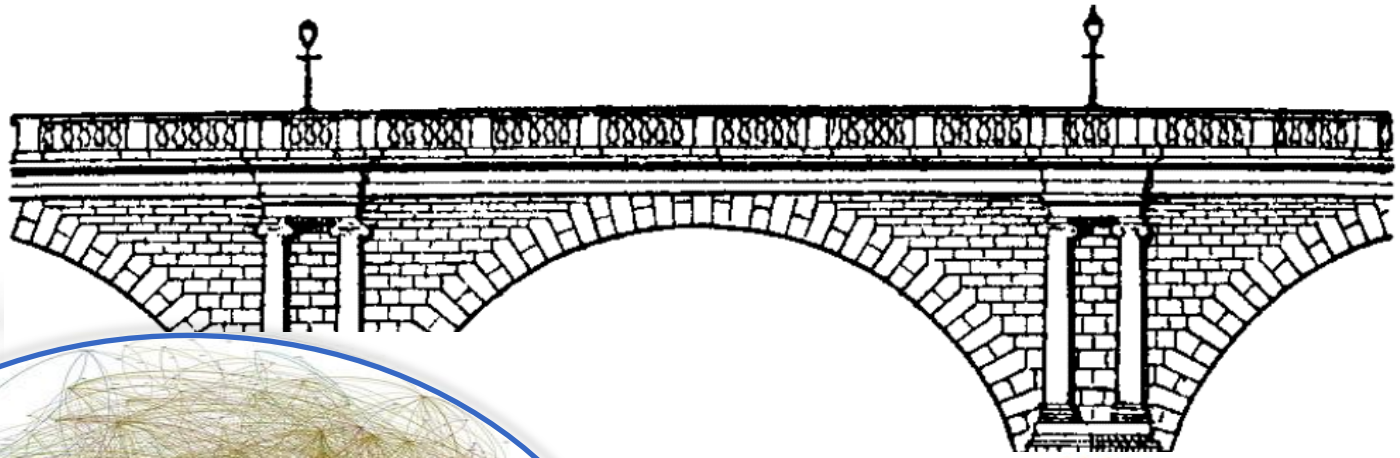
- A significant percentage of information needs are recall-oriented and this justifies the need for exploratory search services
- Faceted Search/Exploration is an effective model for exploratory search (now de facto standard)
- There are many variation of FE according to: structural complexity of the information space, configurability, supported state space
- We have seen a model for FE that captures the majority of datasets that exist

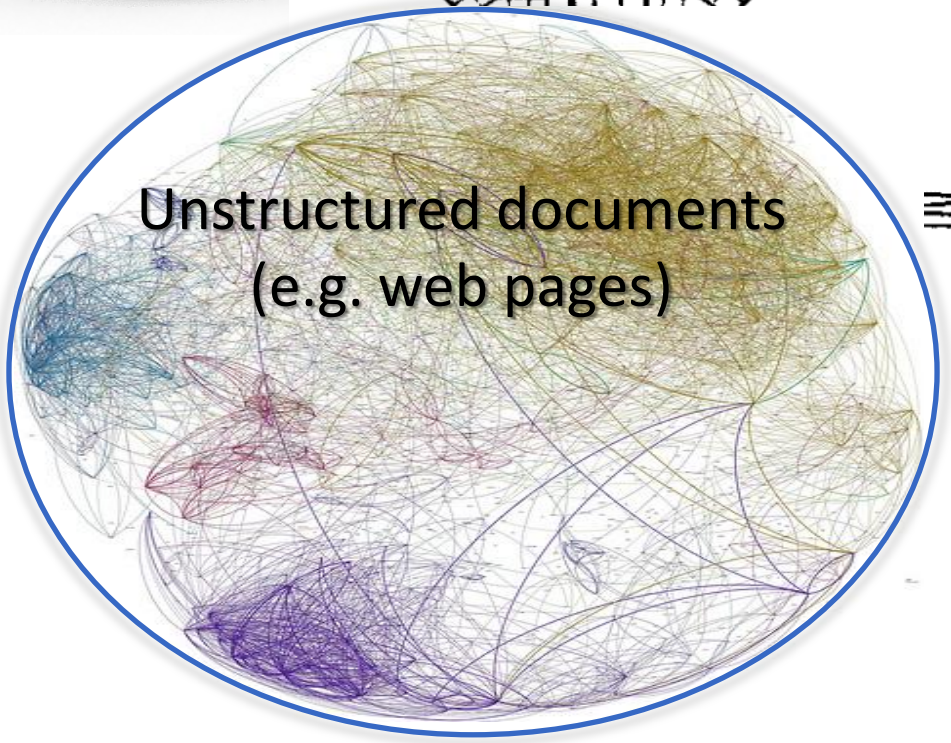
3. Bridging the Web of Documents with the Web of Data at Search time

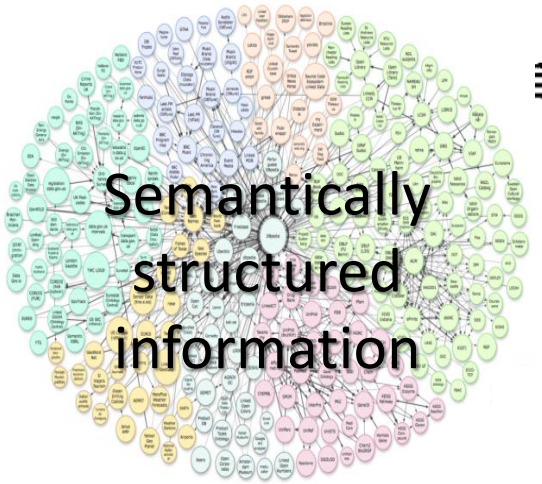
Possible Ways

Focus on doing this at *Search Time*

An interesting question





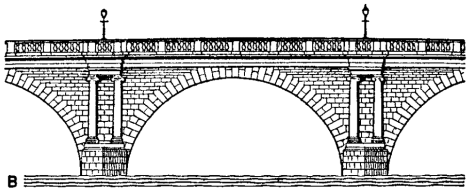


This question is instance of a Bigger Question

- How to integrate the results and/or foster collaboration between different communities

IR Community

Database Community



NLP Community

AI Community

Bridging the Web of Documents with the Web of Data

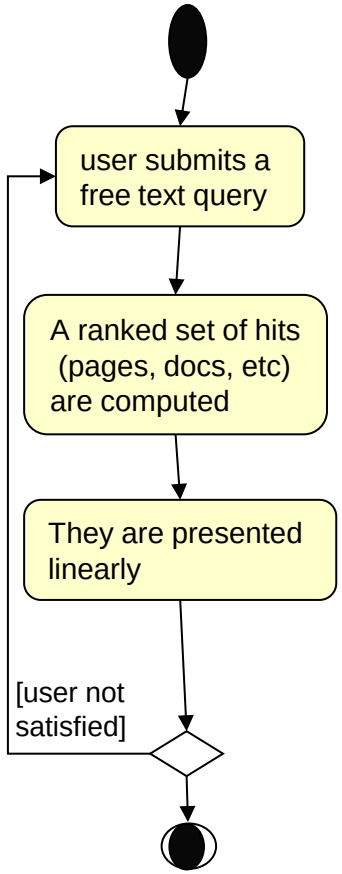
- Why
  - Both exist (why not exploit both?)
- When and Where
  - There is plethora of options. Let's consider the direction: How the searching over the Web of Documents can benefit from the existence of the Web of Data. A few options follow:
    - At indexing time
      - E.g. for the disambiguation of words
    - At query formulation time
      - E.g. autocompletion/query expansion/term_recommendation services that exploit the web of data, the various vocabularies, ontologies, etc.
    - At query evaluation time
      - Exploitation of how words/concepts are connected in the ranking formula
    - After query evaluation
      - For semantically post-processing of search results
    - …

Our Focus in this Talk

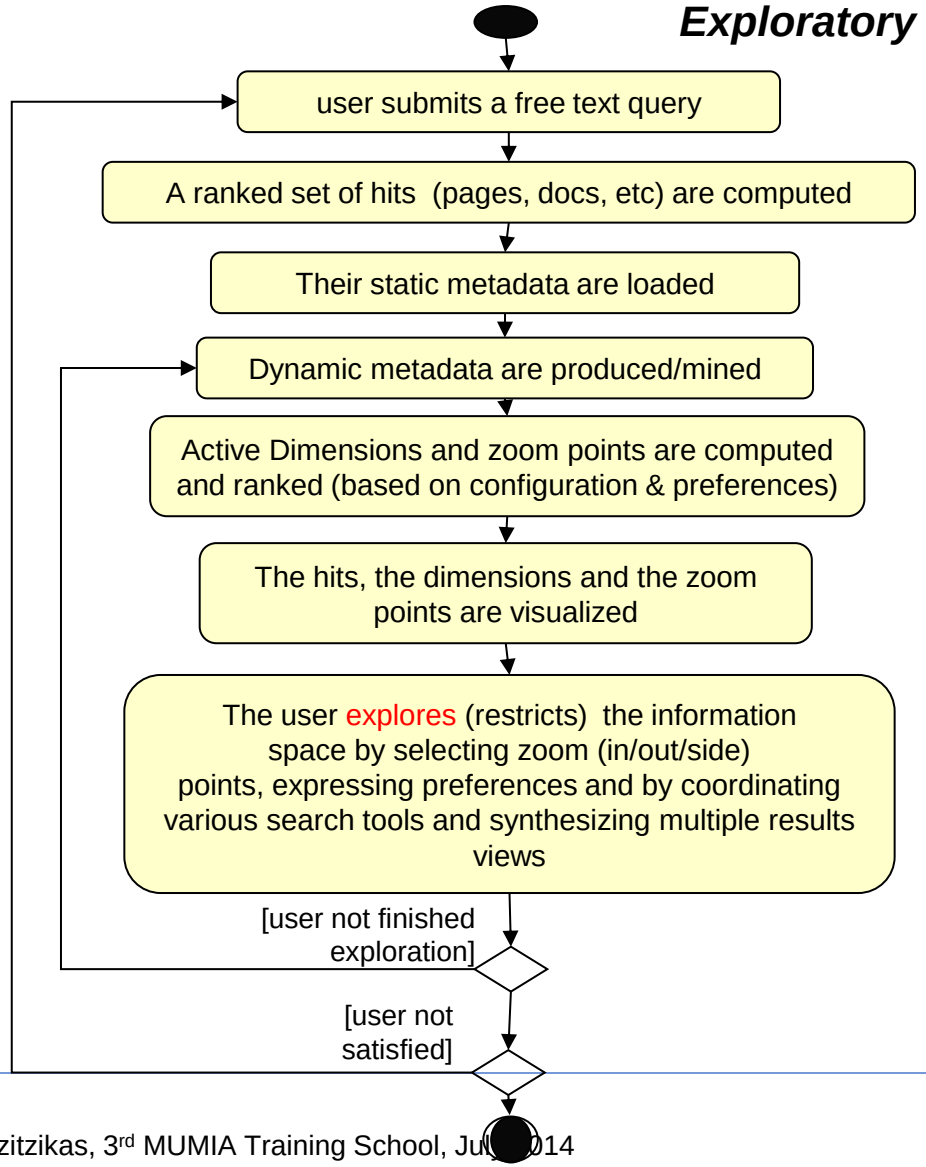
- *Don't change the way users search for information (keyword queries in Web Search Engines).*
- *Try to exploit LOD to offer value-adding services*

A Process for Exploratory Search

Web searching today



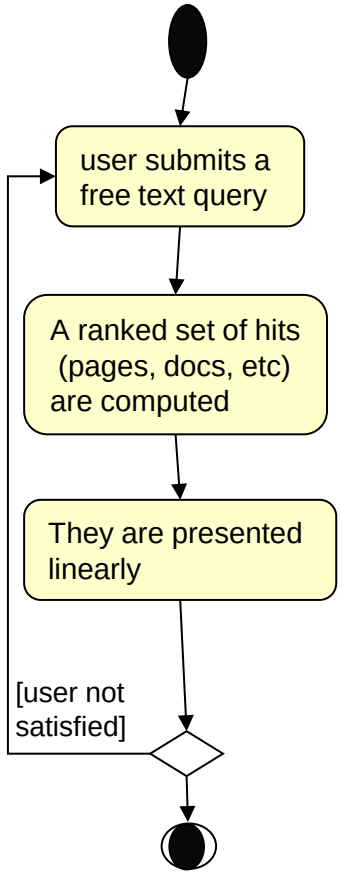
Exploratory Searching



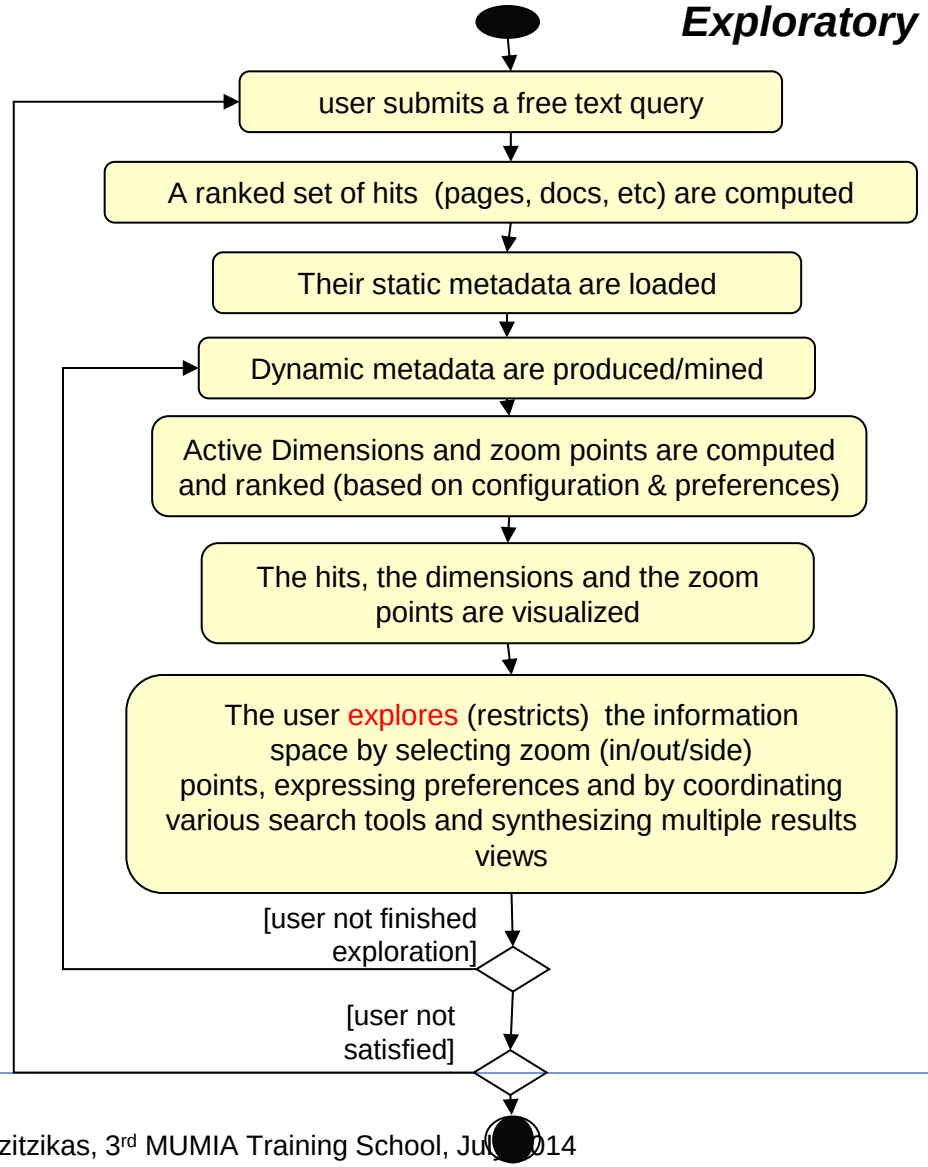
84

A Process for Exploratory Search

Web searching today



Exploratory Searching



85

BENEFITS (OF THIS PROCESS)

- Does not change the way users search for information
- Can be applied over existing search systems/interfaces
- Provides overviews of the results (not only of the top hits)
- By clicking on a facet term (metadata value, mined entity, etc) the user can see the related hits even if they are low ranked
- Allows restricting the answer gradually

Next we will see various applications of this process

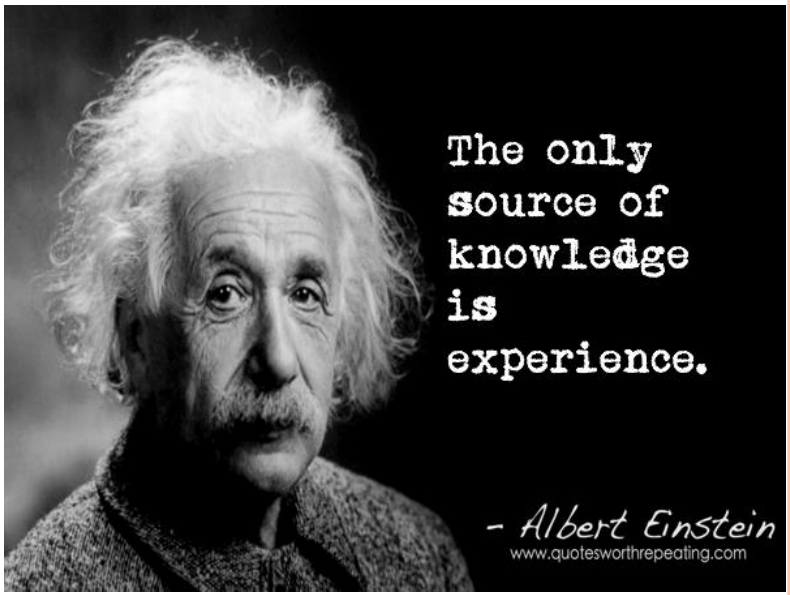
(these applications concern Web Search, Exploratory Search and LOD)

4. Case studies

Case Studies Outline

- As Case Studies I will present you a kind of "story" organized in 9 milestones

- These 9 milestones correspond to activities of ISL (Information Systems Laboratory) of FORTH-ICS.
  - They correspond to the period 2009-2014
  - They are presented in (almost) chronological order.

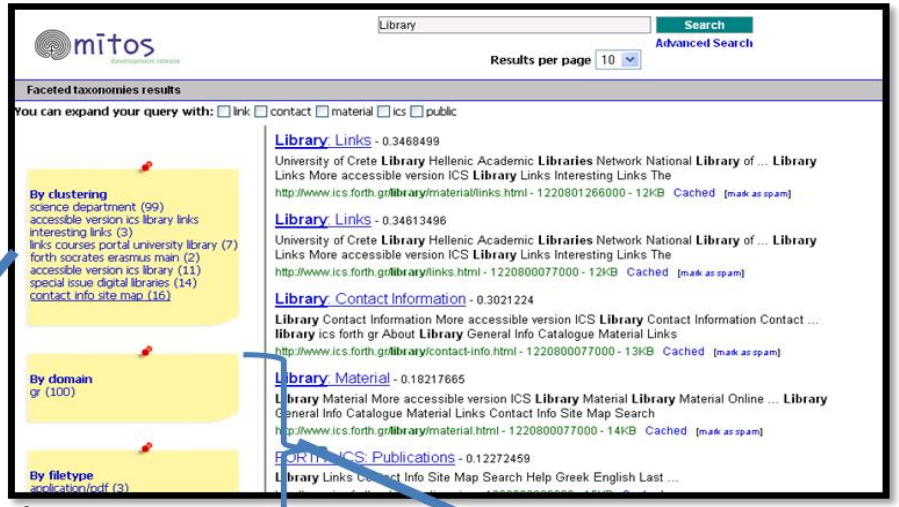


MILESTONE 1. THE MITOS WSE (2009)

mitos development release

- MITOS is WSE build from scratch and collaboratively with the students of the Computer Science Department of the University of Crete.
- Apart from the classical WSE functionality, Mitos offers faceted search over the results of the submitted queries.
  - It supports facets corresponding to metadata attributes of the web pages (static metadata), as well as facets corresponding to the outcome of snippet-based clustering algorithms (a kind of dynamic metadata).
  - The user can then restrict his/her focus gradually, by interacting with the resulting multidimensional structure through simple clicks.

THE MITOS WSE (2009)



Dimension based on dynamic (query-dependent) metadata (of the top ranked hits)

By clustering
- architecture (6)
- contact (10)
- content (10)
- copyright notice csd (13)
- course content english (7)
- csd (14)
- department (8)
- forth (38)
- health telematics network (6)
- ics (39)
- information (11)
- network (8)
- physical (5)
- science (22)
- ITE τεχνικές αναφορές (22)

REST (3981)

Dimensions based on static metadata

By domain
- gr (4067)

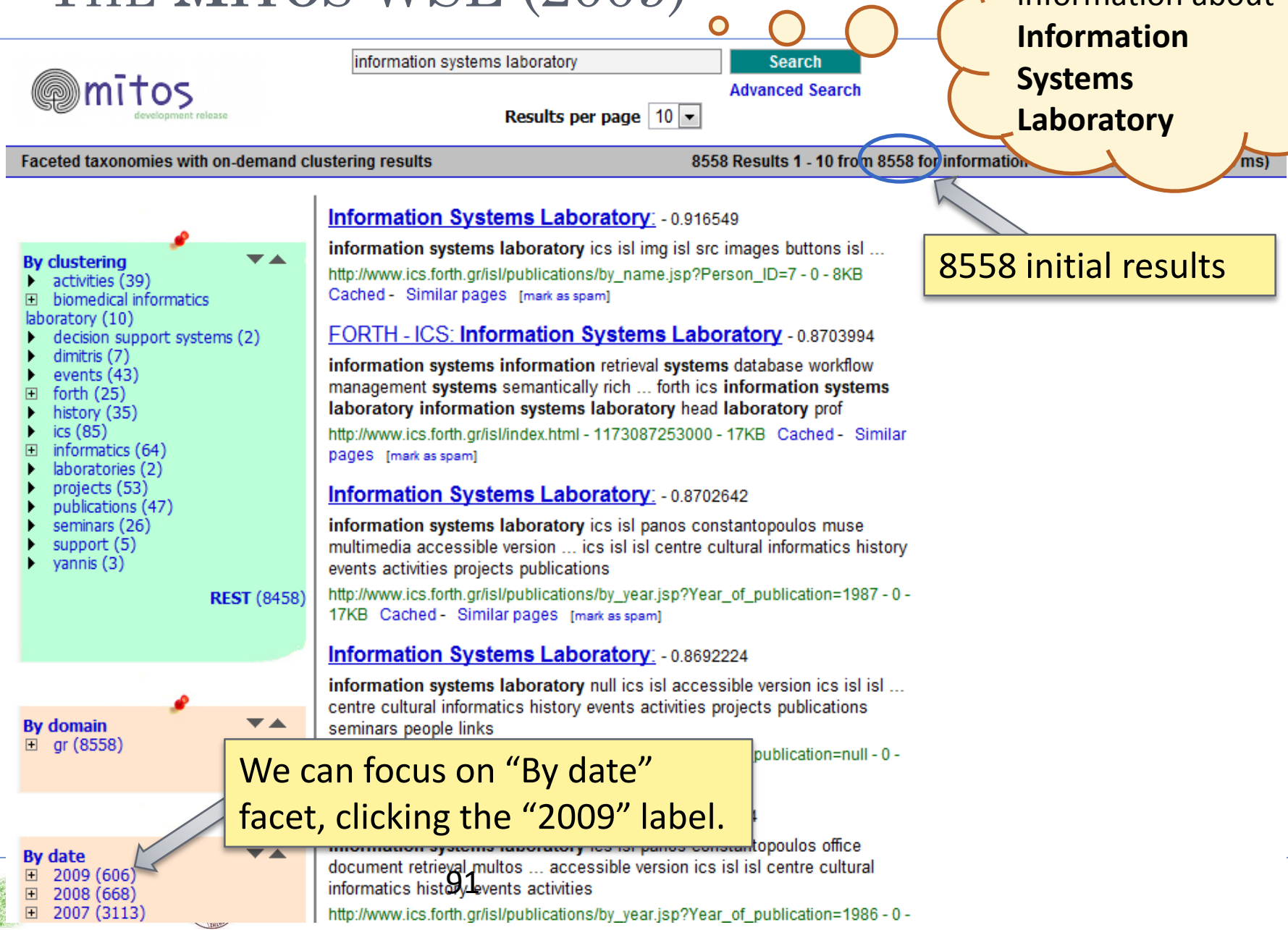
By date
- 2008 (479)
- 2007 (694)
- 2006 (1340)
- 2005 (184)
- 2004 (106)
- 2003 (82)
- 2002 (88)
- 2001 (28)
- 2000 (13)
- 1999 (4)
- 1998 (6)
- 1997 (1)
- Unknown (1042)

By filetype
- application/msword (16)
- application/pdf (1476)
- application/vnd.ms-powerpoint (29)
- text/html (2546)

By language
- Any (UTF-8) (18)
- Greek (1209)
- Latin-1 (Europe, Latin America, Caribbean, Canada, Africa) (944)
- Latin-2 (Central and Eastern Europe) (4)
- Unknown (1892)

N. Manolis and Y. Tzitzikas (ESWC'11)

THE MITOS WSE (2009)

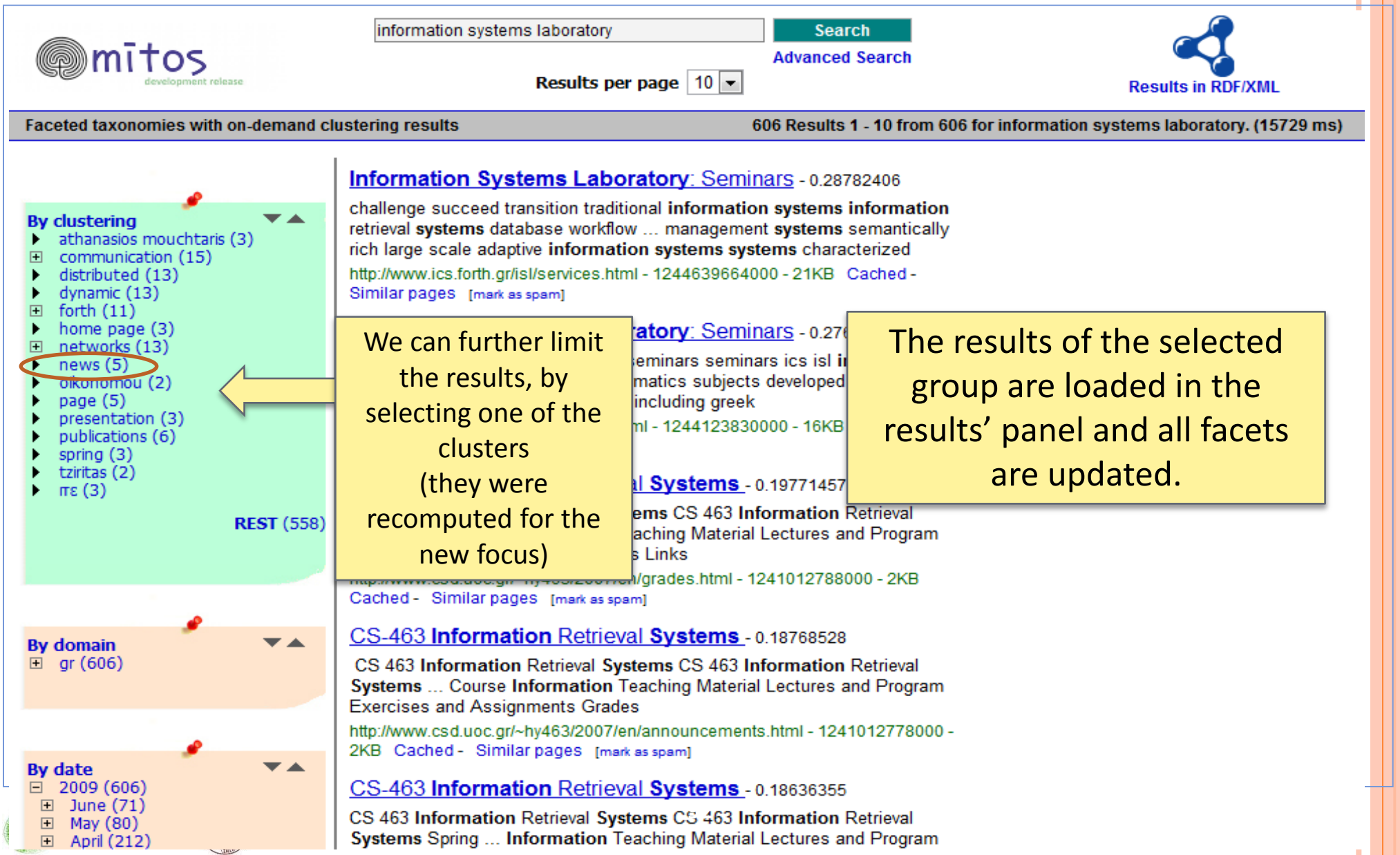


A user wants to get information about **Information Systems Laboratory**

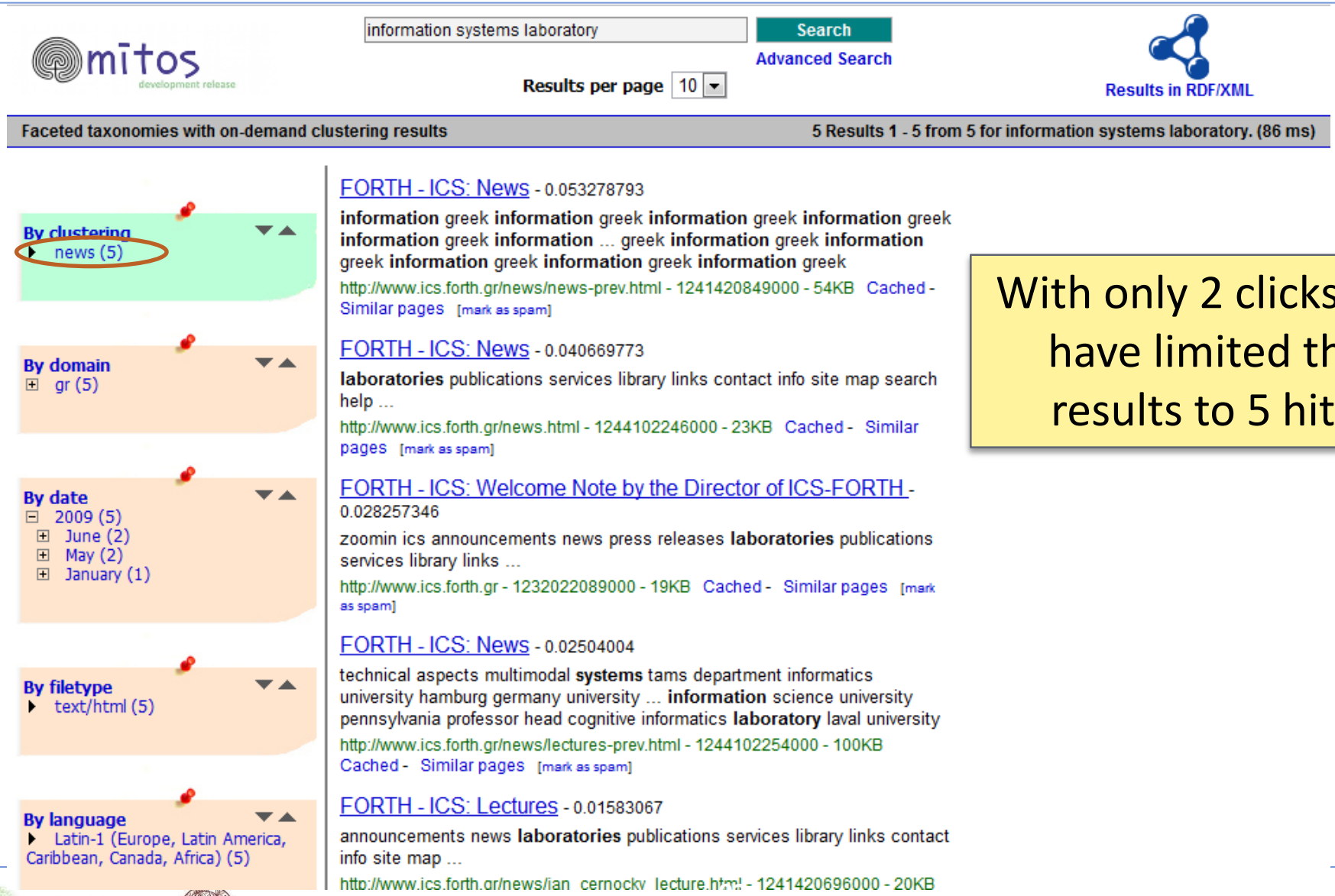
8558 initial results

We can focus on "By date" facet, clicking the "2009" label.

(CONT.)


mītos
development release
information systems laboratory
Search
Advanced Search
Results per page 10
Results in RDF/XML
Faceted taxonomies with on-demand clustering results
606 Results 1 - 10 from 606 for information systems laboratory. (15729 ms)
By clustering
athanasios mouchtaris (3)
communication (15)
distributed (13)
dynamic (13)
forth (11)
home page (3)
networks (13)
news (5)
oikonomou (2)
page (5)
presentation (3)
publications (6)
spring (3)
tziritas (2)
πε (3)
REST (558)
By domain
gr (606)
By date
2009 (606)
June (71)
May (80)
April (212)
Information Systems Laboratory: Seminars - 0.28782406
challenge succeed transition traditional information systems information retrieval systems database workflow ... management systems semantically rich large scale adaptive information systems systems characterized
http://www.ics.forth.gr/isl/services.html - 1244639664000 - 21KB Cached - Similar pages [mark as spam]
We can further limit the results, by selecting one of the clusters (they were recomputed for the new focus)
The results of the selected group are loaded in the results' panel and all facets are updated.
CS-463 Information Retrieval Systems - 0.18768528
CS 463 Information Retrieval Systems CS 463 Information Retrieval Systems ... Course Information Teaching Material Lectures and Program Exercises and Assignments Grades
http://www.csd.uoc.gr/~hy463/2007/en/announcements.html - 1241012778000 - 2KB Cached - Similar pages [mark as spam]
CS-463 Information Retrieval Systems - 0.18636355
CS 463 Information Retrieval Systems CS 463 Information Retrieval Systems Spring ... Information Teaching Material Lectures and Program

(CONT.)



With only 2 clicks, we have limited the results to 5 hits.

CONT.

- **Evaluation** with Users (main results) :
  - Faceted search, combining dynamically and statically mined metadata
    - lead to much **improved task completeness** with much **less user interactions**
    - was **more preferred** by the users (advanced and plain ones) and lead **to greater satisfaction**, than plain clustering or faceted interfaces
- Most Important Related Publications
  - [ECDL'09] P. Papadakos, S. Kopidaki, N. Armenatzoglou and Y. Tzitzikas. Exploratory Web Searching with Dynamic Taxonomies and Results Clustering. In ECDL 2009
  - [WISE'09] S. Kopidaki, P. Papadakos, and Y. Tzitzikas. STC+ and NM-STC: Two novel online results clustering methods for web searching. In WISE 2009
  - [J. KAIS 2012] P.Papadakos, S.Kopidaki, Nikos Armenatzoglou and Y. Tzitzikas On exploiting Static and Dynamically mined Metadata for Exploratory Web Searching , KAIS Journal, 2012

MILESTONE 2. DURING TYPING?

- Then we questioned ourselves:
  - ***why not offering this functionality during query typing, i.e. a kind of richer autocompletion service?***
- This resulted to what we called **Instant Overview Search (IOS)**.
- The idea:
  - For the frequent queries, pre-compute and store not only the first page of results, but also the analysis of these hits
- Technical challenge
  - Since the amount of information that has to be stored for each query is higher (and obviously does not fit in main memory) we devised a partitioned trie-like index for efficiency (plus a dedicated cache)

IOS (INSTANCE OVERVIEW SEARCH), 2011-2012

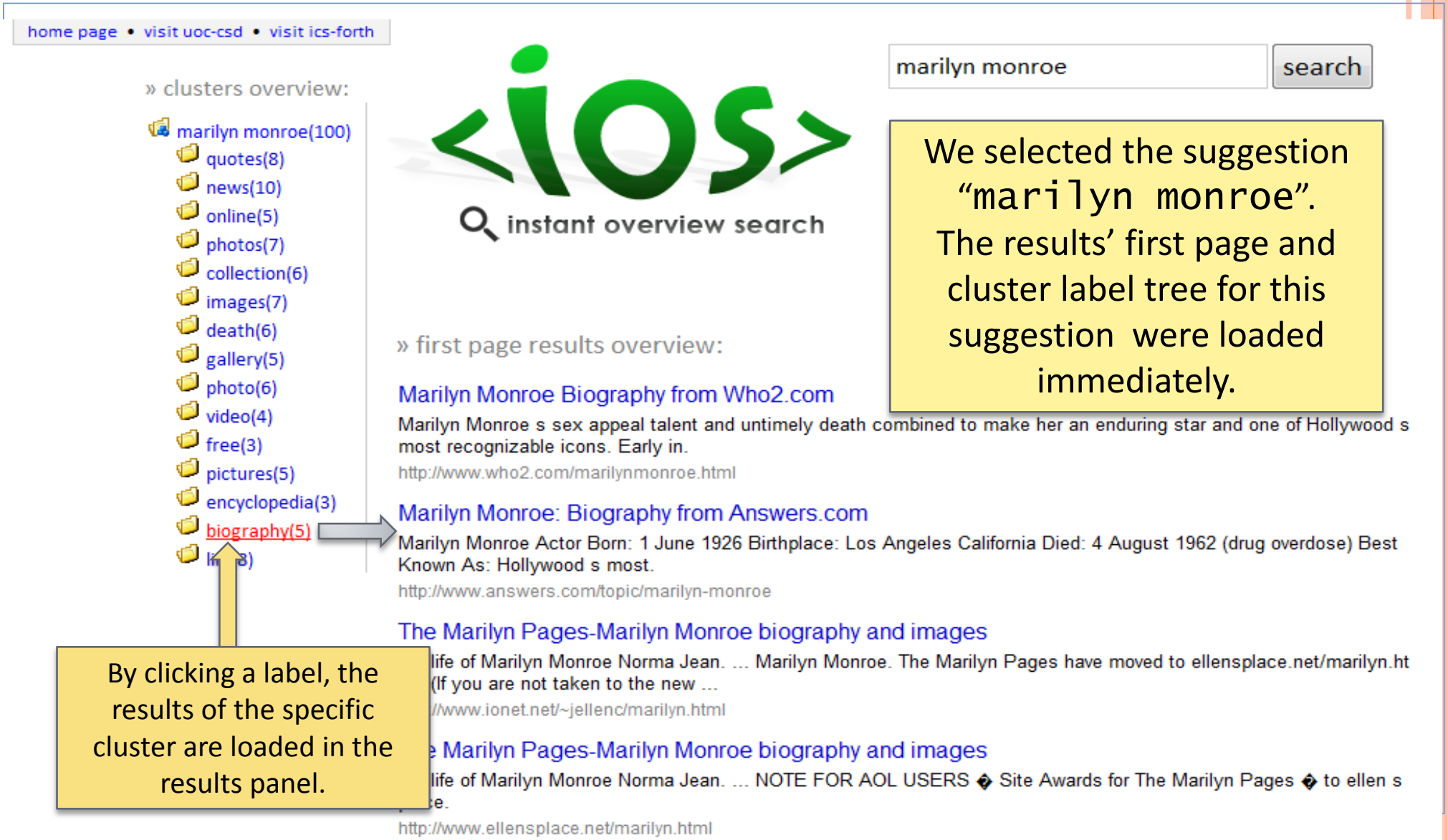
This is Pavlos Fafalios
(I supervise his PhD)

We want to find information about the life of Marilyn Monroe
(and probably its connection to Pavlos?)
However, we are not sure for the spelling of her name.
So, we start typing "`mari`".

(CONT.)

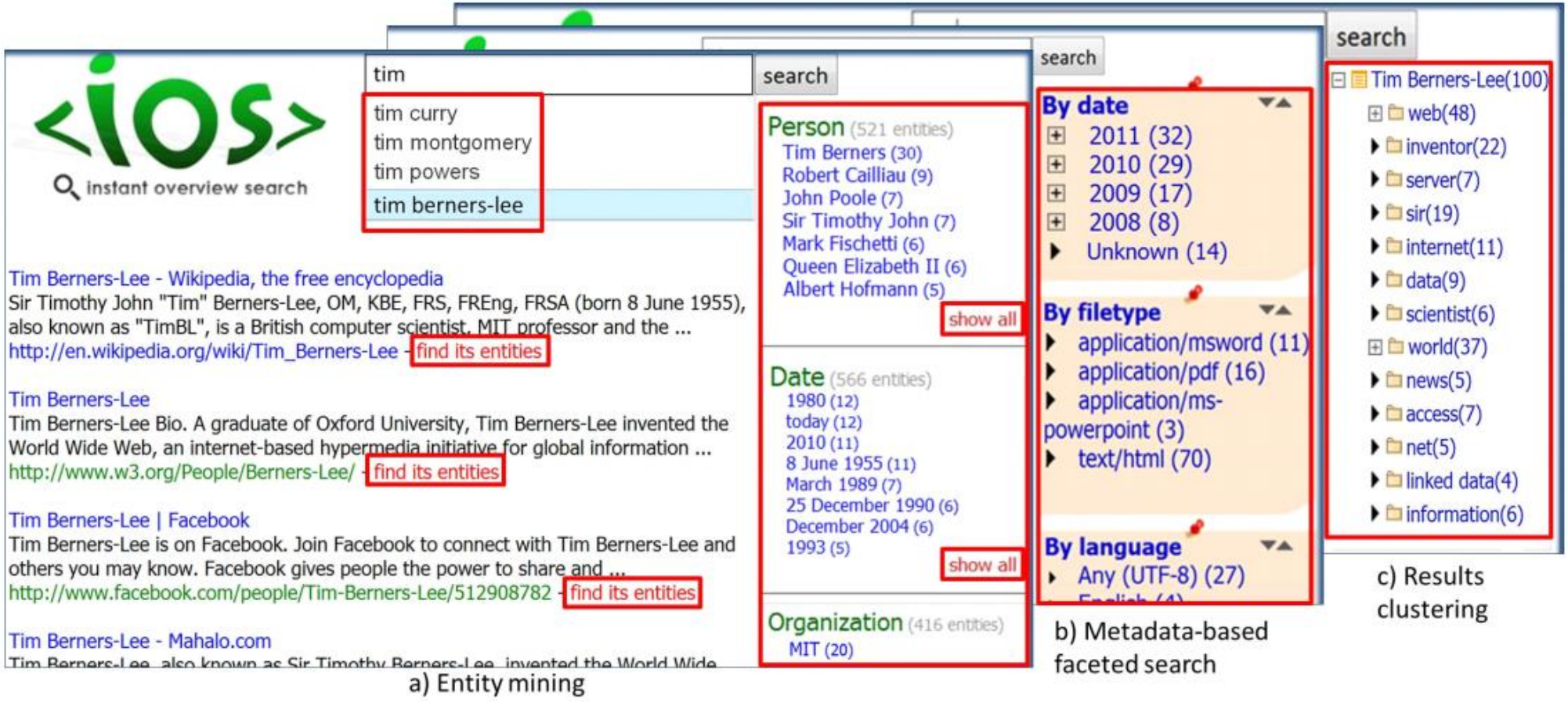

home page • visit uoc-csd • visit ics-forth
» clusters overview:
marilyn(99)
monroe(33)
quotes(4)
biography(3)
manson(8)
home(8)
free(7)
encyclopedia(3)
quotes(5)
photo(5)
collection(5)
news(5)
encyclopedia(4)
biography(4)
new(10)
york(4)
gavin rossdale(3)
images(5)
andy(3)
online(3)
instant overview search
mari
search
marilyn
marilyn monroe
mario games
• SET (default)
• PET
• STIE
• PTIE
» over linked open data
List of query's suggestions.
First page of results of the top suggestion "marilyn"
» first page results overview:
Marilyn (singer) - Wikipedia, the free encyclopedia - 0
Peter Robinson (born 3 November 1962), better known as Marilyn, is a British pop singer who achieved international fa me in the 1980s with his hit song ...Early life - Blitz years - Career - Recent activity
en.wikipedia.org/wiki/Marilyn_(singer)
Marilyn Monroe - Wikipedia, the free encyclopedia - 1
Marilyn Monroe born Norma Jeane Mortenson, but baptized Norma Jea Miller - Somethings Got to Give - Some Like It Hot
en.wikipedia.org/wiki/Marilyn_Monroe
Marilyn (hill) - Wikipedia, the free encyclopedia - 2
A Marilyn is a mountain or hill in the United Kingdom, Ireland or Isle of ...
en.wikipedia.org/wiki/Marilyn_(hill)
Agency - 3
model agency in Paris. Marilyn Agency est la plus importante agence de mannequin Γ
We can continue typing the query. Instantly new suggestions are shown
Cluster Label Tree of the top suggestion "marilyn"

(CONT.)

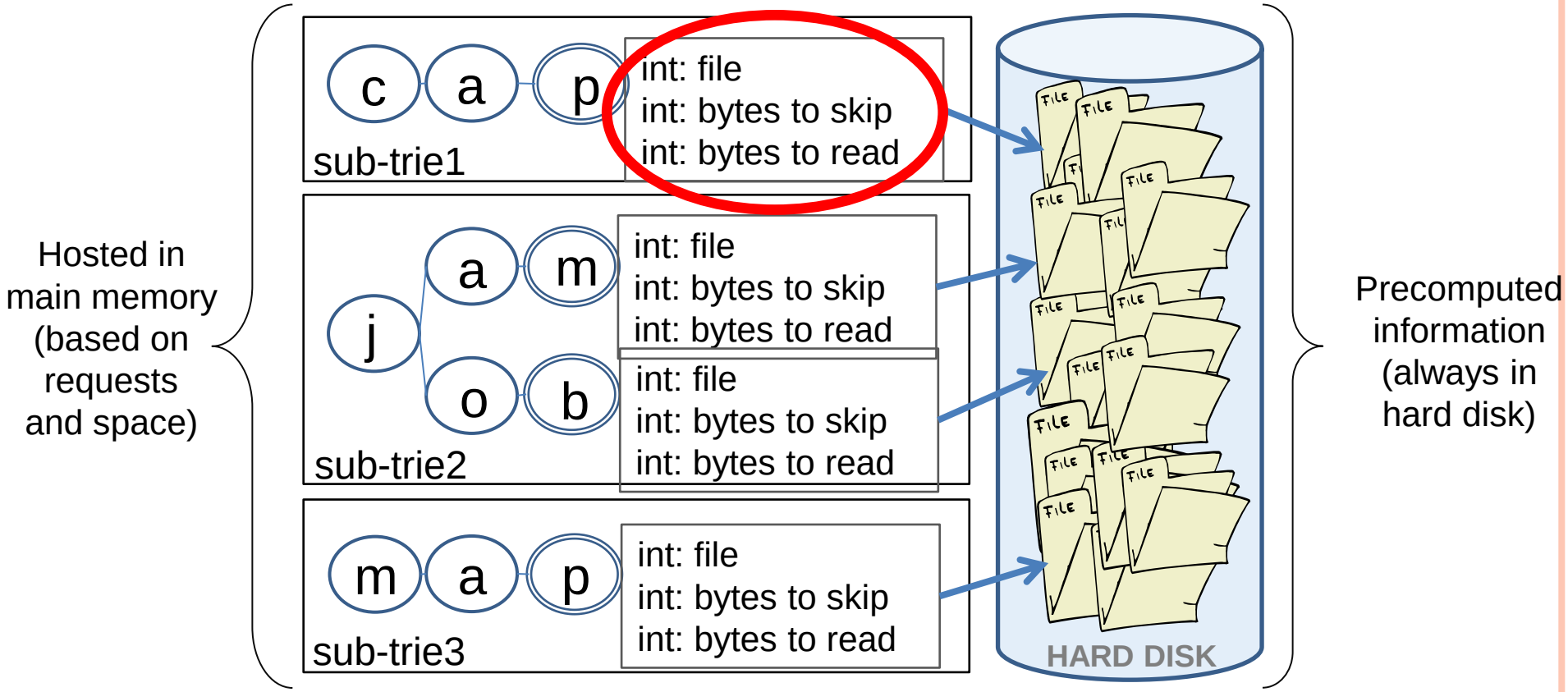

home page • visit uoc-csd • visit ics-forth
» clusters overview:
marilyn monroe(100)
quotes(8)
news(10)
online(5)
photos(7)
collection(6)
images(7)
death(6)
gallery(5)
photo(6)
video(4)
free(3)
pictures(5)
encyclopedia(3)
biography(5)
instant overview search
marilyn monroe
search
We selected the suggestion "marilyn monroe". The results' first page and cluster label tree for this suggestion were loaded immediately.
» first page results overview:
Marilyn Monroe Biography from Who2.com
Marilyn Monroe s sex appeal talent and untimely death combined to make her an enduring star and one of Hollywood s most recognizable icons. Early in.
http://www.who2.com/marilynmonroe.html
Marilyn Monroe: Biography from Answers.com
Marilyn Monroe Actor Born: 1 June 1926 Birthplace: Los Angeles California Died: 4 August 1962 (drug overdose) Best Known As: Hollywood s most.
http://www.answers.com/topic/marilyn-monroe
The Marilyn Pages-Marilyn Monroe biography and images
http://www.ionet.net/~jellenc/marilyn.html
http://www.ellensplace.net/marilyn.html
By clicking a label, the results of the specific cluster are loaded in the results panel.

IOS (INSTANCE OVERVIEW SEARCH), 2011-2012

We can exploit this technique for any king of pre-processing of search results (e.g. metadata-based faceted search, snippet-based clustering, entity mining, etc)



IOS INDEXES



Average Retrieval Time ≈ **135ms**

Experiments *over a server running on a* ***modest personal computer****, with a synthetic query log of* ***1 million*** *distinct queries and synthetic precomputed information of* ***1 Terabyte***

CONT

- Key results
  - A partitioned trie-based index structure that can efficiently support recommendations for millions of distinct queries even with modest hardware
    - One can provide instant access to large amount of data, utilizing the existing resources, without requiring more hardware
  - A hybrid caching policy (70% static and 30% dynamic) seems to be the more appropriate choice yielding a throughput increment of around 80% and a 25% speedup
- Demo
  - http://www.ics.forth.gr/isl/ios
    - Select the system "Instant Entity Mining + Clustering (over Bing)"
- Related Publications
  - [WISE'11] P. Fafalios and Y. Tzitzikas, Exploiting Available Memory and Disk for Scalable Instant Overview Search, 12th International Conference on Web Information System Engineering (WISE 2011), Sydney, Australia, October 2011
  - [WWW'12] P. Fafalios, I. Kitsos and Y. Tzitzikas, Scalable, Flexible and Generic Instant Overview Search, 21st International Conference on World Wide Web, (WWW 2012), Demo Paper, Lyon, France, April 2012

MILESTONE 3. ENTITY MINING AND LOD?

- Then we questioned ourselves:
  - ***why not exploiting LOD in the context of entity mining of the search results?***
- Motivation
  - LOD contains plenty of information about Named Entities (their names, attributes, relationships with other entities, etc)
- Output
  - IOS Entity Mining
    - LOD is used as source for Named Entity Recognition
    - LOD is used for providing more information about the identified entities

IOS Entity Mining (2012)

http://www.ics.forth.gr/isl/ios

- Automatically connects knowledge with documents at query time
- No preprocessing
- No indexing

CONT

- Exploitation for restricting the focus

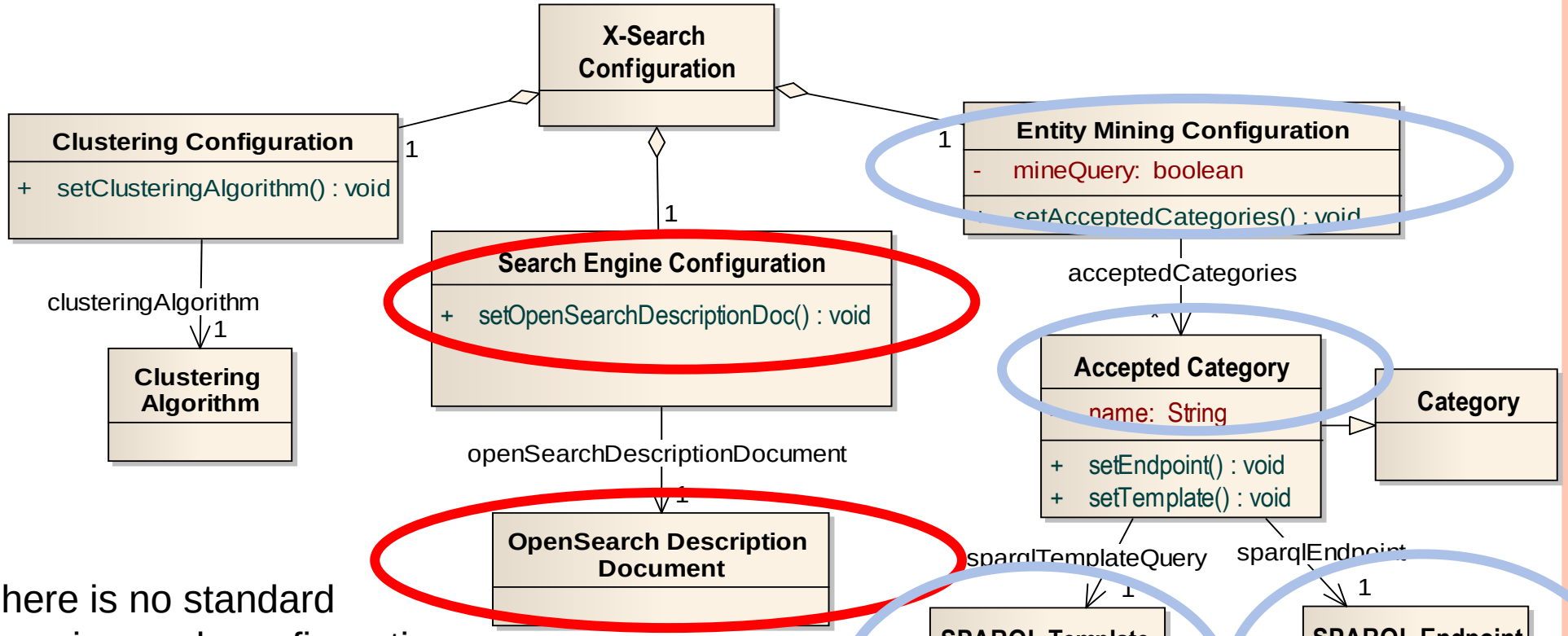
CONT.

- Some results
  - Real-time NEM over snippets is feasible and yields about 1.2 entities per snippet
  - NEM over contents is more time consuming, but mines much more entities
  - String similarity between the query and the entity name does not improve entity ranking (in our setting)
  - The top-10 entities derived from snippet mining are quite different from those derived from contents mining (< 30% Jaccard similarity)
- Related Publications
  - [IRFC'12] P. Fafalios, I. Kitsos, Y. Marketakis, C. Baldassarre, M. Salampasis and Y. Tzitzikas, *Web Searching with Entity Mining at Query Time,* 5th Information Retrieval Facility Conference (IRFC 2012), Vienna, Austria, July 2012

MILESTONE 4. CONFIGURABILITY (AND LOD)

- Then we questioned ourselves:
  - ***why not allowing the user to configure himself the entities of interest by exploiting LOD (again in the context of entity mining of the search results)?***
- Outcome
  - X-ENS (e**X**plore **EN**tities in **S**earch)
- Related Publications
  - [SIGIR'13] P. Fafalios and Y. Tzitzikas, X-ENS: Semantic Enrichment of Web Search Results at Real-Time, 36th International ACM SIGIR Conference, Demo Paper, Dublin, Ireland, 28 July - 1 August 2013

XSEARCH-CONFIGURABILITY: THE CONCEPTUAL MODEL



There is no standard covering such configurations. We have defined our own Configuration format

X-ENS (SIGIR 2013)

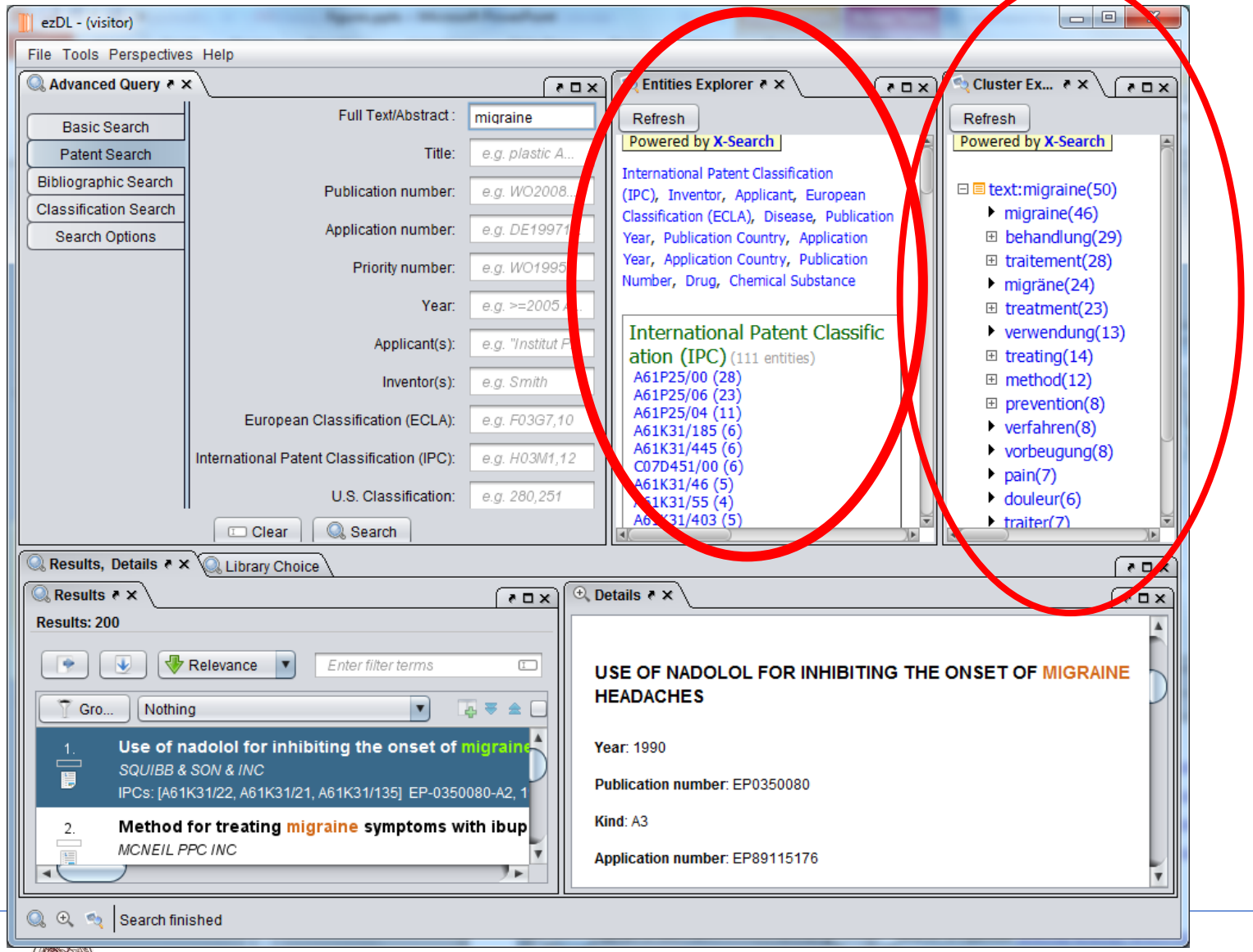
http://139.91.183.72/x-ens

MILESTONE 5. PROFESSIONAL SEARCH SYSTEMS?

- Then we questioned ourselves:
  - ***why not applying and testing this in the context of a professional search system?***
- Outcome
  - Application in **patent search**. Missing relevant documents is unacceptable in patent search *(recall oriented search procedure)*. Retrieval of all relevant documents is usually necessary
    - Patents contain plenty of named entities of various kinds
      - *Companies, Countries, Persons, Product types, Laws, etc*
  - Inclusion of **PerFedPat** System
    - In collaboration with Mike Salampasis

PERFEDPAT

http://www.perfedpat.eu/



PERFEDPAT (CONT)

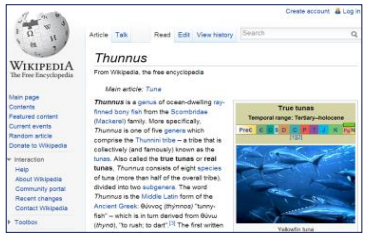
- The proposed functionality:
  - offers a tight integration of different search tools with the main retrieval engine,
  - connects the search results (i.e. patents) with data and knowledge,
  - can be exploited by any patent search system (i.e. it acts as a service over a ranked list of results)
  - The time that we have to pay is proportional to the number of the top results that we want to "explore" (≈ 1.5 sec / 100 results)
- Related Publications
  - P. Fafalios, M. Salampasis and Y. Tzitzikas, Exploratory Patent Search with Faceted Search and Configurable Entity Mining, 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with ECIR'13, Moscow, Russia, March 2013

MILESTONE 6

APPLYING IN THE CONTEXT OF AN INFRASTRUCTURE

- Then we questioned ourselves:
  - ***why not applying this in another domain of professional search in the context of a real and operating EU research infrastructure?***
- Outcome
  - X-Search in the context of the ongoing iMarine Research Infrastructure project

XSEARCH:
SEMANTIC POST-PROCESSING OF SEARCH RESULTS



http://www.i-marine.eu/

i-marine
Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources

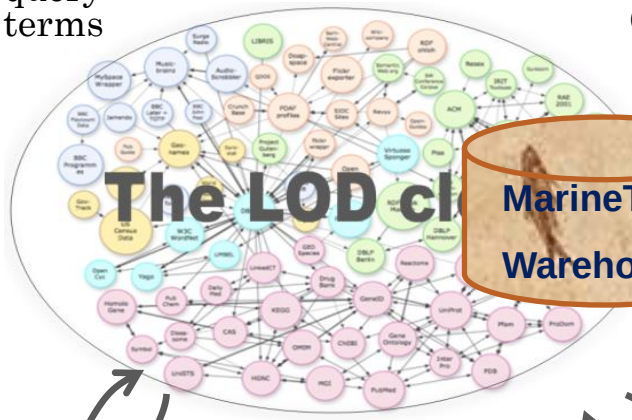
web browsing

contents

gCube BE Google bing OpenSearch

query terms

(top-L) results (+ metadata)



MarineTLO Warehouse

Entity Mining

entities / contents

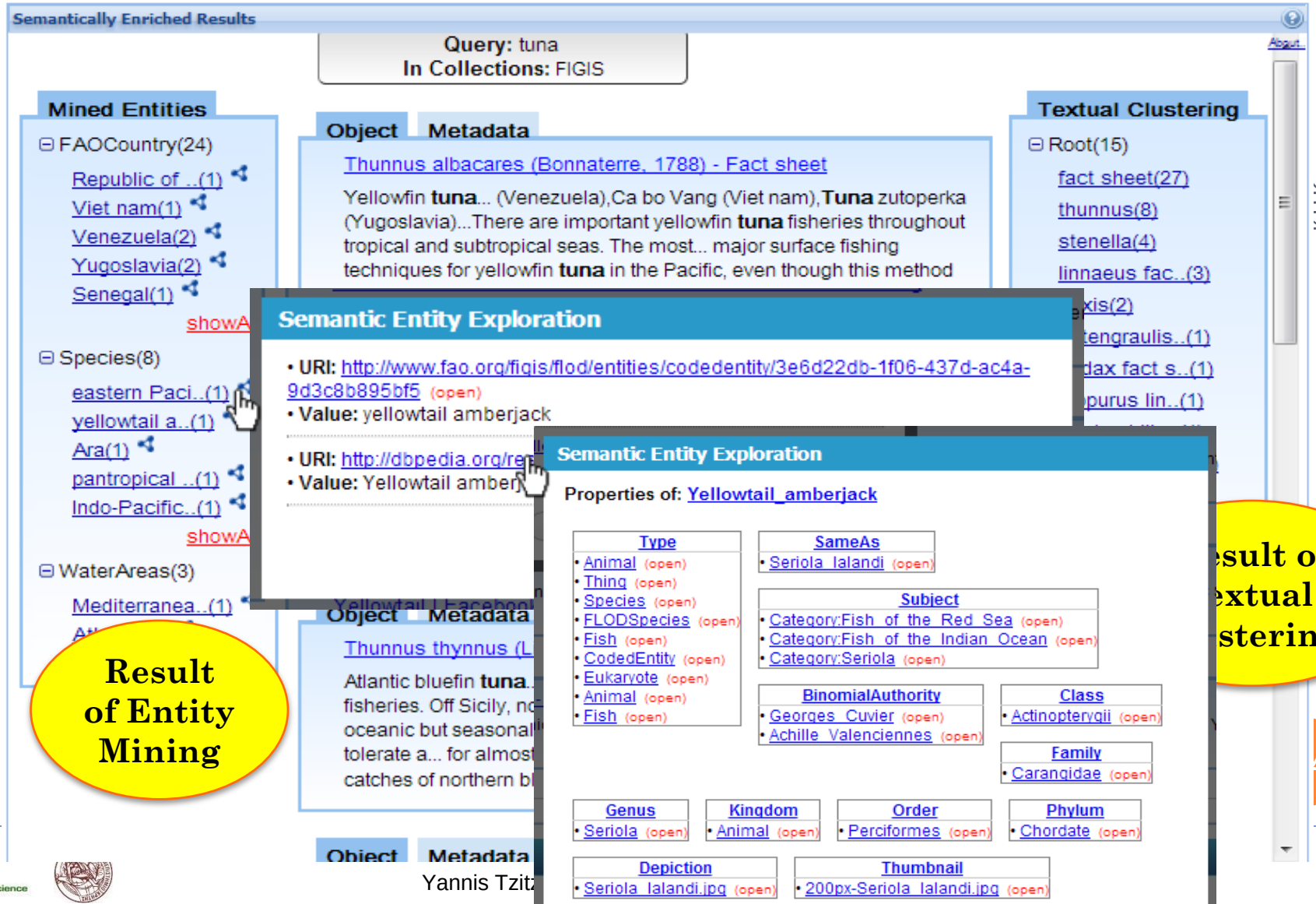
Visualization/Interaction
(faceted search, entity exploration, annotation, top-k graphs, etc.)

semantic data

Semantic Analysis

- Grouping,
- Ranking
- Retrieving more properties

Example: X-Search deployed in an Operational Research Infrastructure (2012-now)



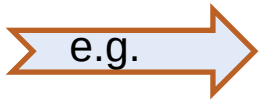
MILESTONE 7 PARALLELIZATION

- Then we questioned ourselves:
  - ***Can we do the same over the full contents of the search hits? Downloading takes time. The processing also is expensive***
- Outcome
  - Investigation of how the task can be partitioned to several machines
  - A MapReduce-based parallelization of downloading and entity mining task for exploiting the resources of cloud
- Key points
  - How to tackle the uncertainty (of hits size) for achieving load balancing and reaching the ideal speedup
  - Two processes for carrying out this task

Levels of Functionality

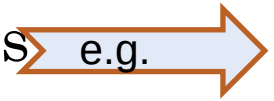
Centralized Process

- ***L0***: Categories + Entities



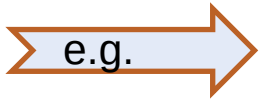
Location
Athens
Greece

- ***L1***: *L0* + Count Info for Entities



Location
Athens (3)
Greece (5)

- ***L2***: *L1* + Rank Entities



Location
Greece (5)
Athens (3)

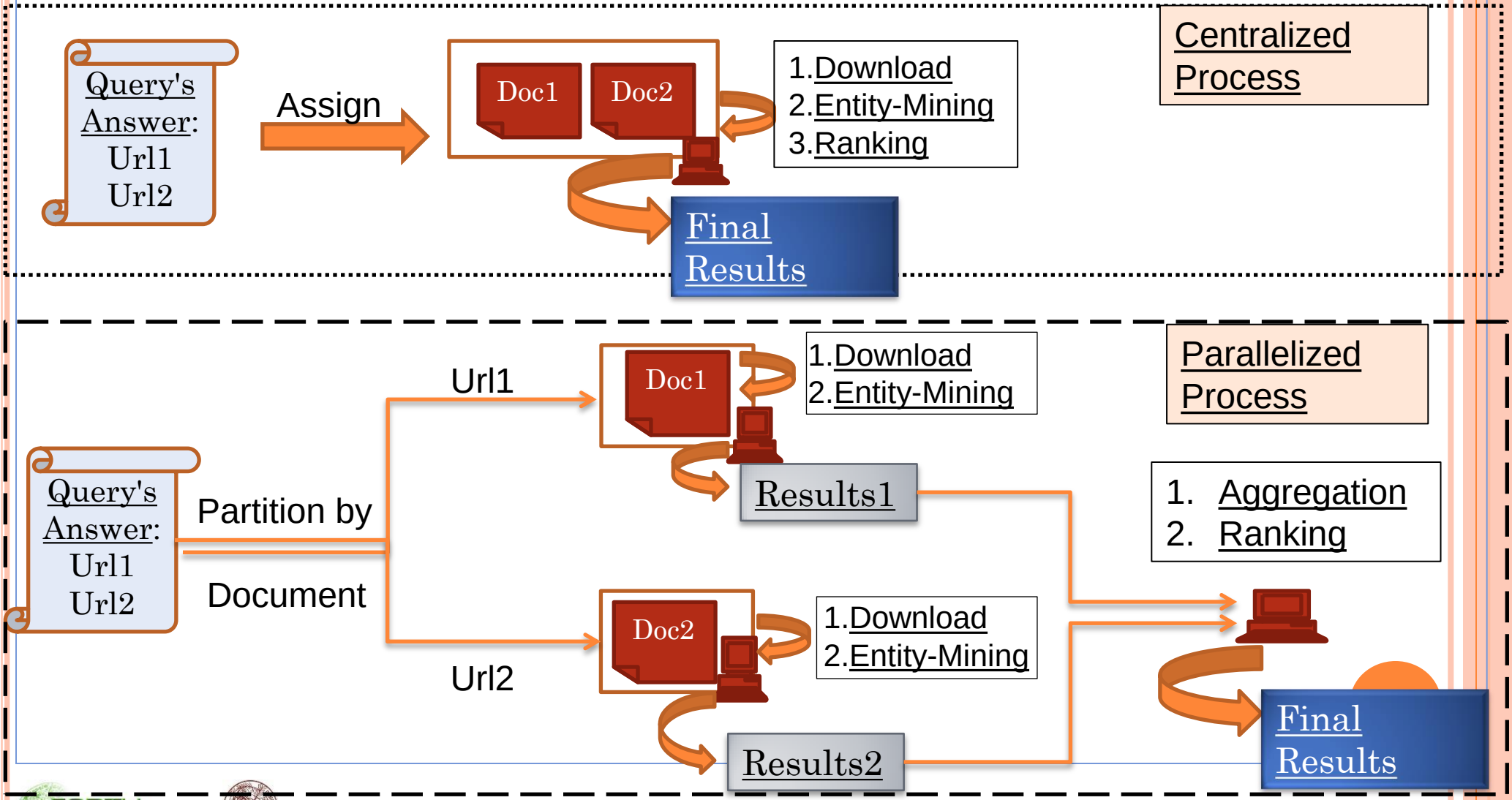
Ranked Entities

- ***L3***: *L2* + Doc List for each entity
- ***L4***: *L3* + Semantic Enrichment

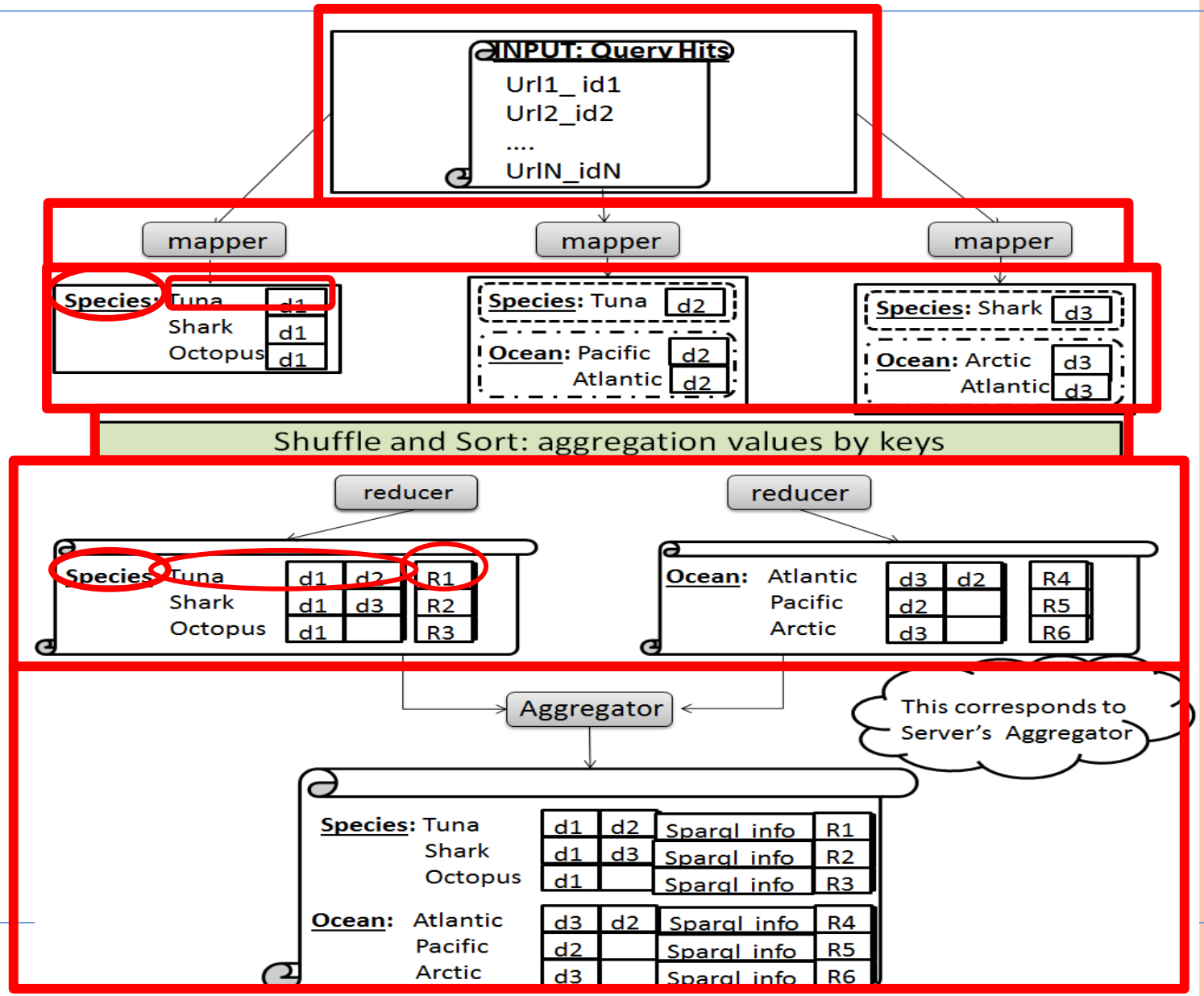
Location
Greece (5) LOD_URL
Athens (3) LOD_URL

THE PARALLELIZATION PROCEDURE

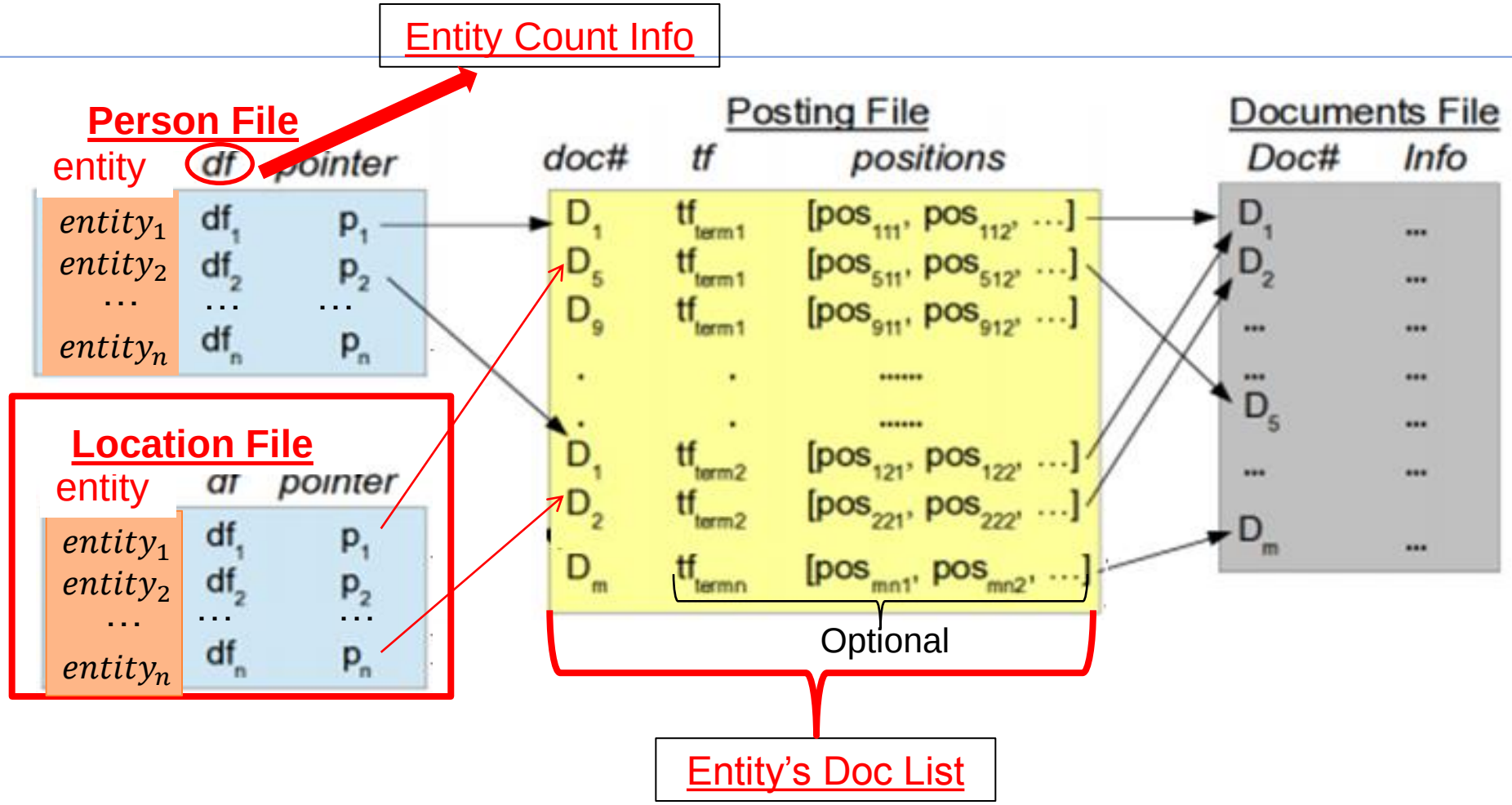
- Main idea: *partition by documents (hit)*



EXAMPLE OF DISTRIBUTED NEM PROCESSING USING MAPREDUCE



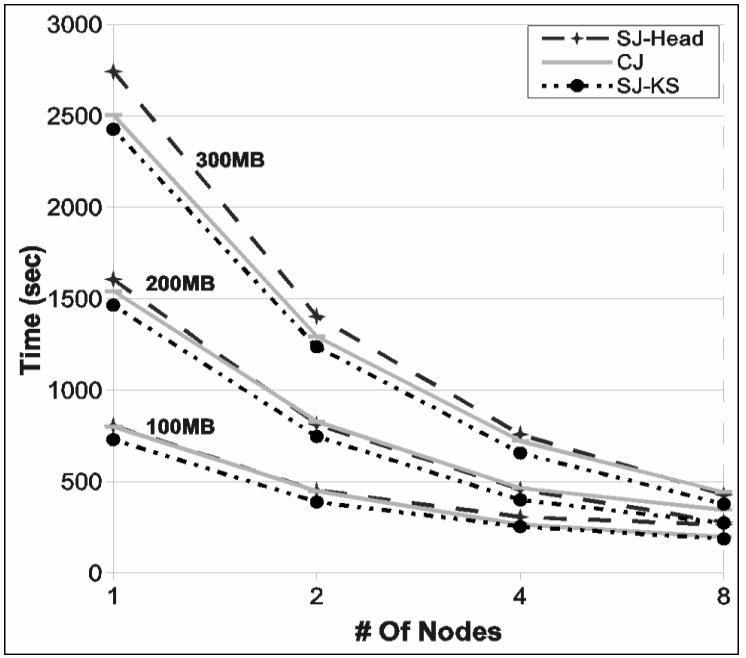
An Analogy to Inverted Files



- Note that our task is much more CPU and memory intensive
- Requires document's downloading

EXPLOITING MORE THAN ONE MACHINES [J. DAPD 2014]

- Two models of evaluation were investigated
- A thorough evaluation of the parameters that affect performance were conducted
- We reached a speedup close to the ideal (according to Amdahl's law)!



- Related Publications
  - I. Kitsos, K. Magoutis and Y. Tzitzikas, Scalable Entity-based Summarization of Web Search Results using MapReduce, Journal on Distributed and Parallel Databases (DAPD), 32(3), 2014


RUBICON

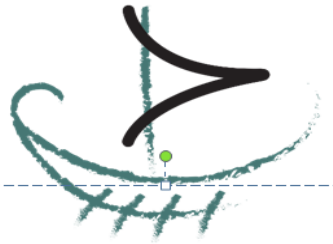

WHAT NEW CAR SHOULD I BUY?
WHY ARE YOU BUYING IT?
TRANSPORT
FUN
WHAT DO YOU CARE ABOUT?
BEING ACTIVE/DOING THINGS
WHAT PEOPLE THINK OF ME
ALL THE MONEY I DON'T HAVE
THE ENVIRONMENT
MY GROWING FAMILY
FORD FOCUS
HONDA FIT
WHAT DO YOU WANT THEM TO THINK?
WHAT IF PEOPLE KNOW YOU'RE BROKE?
FINE, WHATEVER
FORD FIESTA
CHEVY SPARK
NISSAN VERSA NOTE
I'LL DISAPPEAR
HONDA CIVIC
MERCEDES CLS
AUDI A7
LOVE IT
HOPE YOU LIKE DEBT
I SEEM RICH
NO, I REALLY AM RICH
I'M WEIRD
I'M A GANGSTER OR SOMETHING
CHRYSLER 300
BUT I DON'T
MAZDA6
HYUNDAI GENESIS
KIA OPTIMA
DO YOU LIKE THE SKY?
EH, IT'S OKAY
OH YEAH
WHAT KIND OF WEIRD?
ASYMMETRIC
HYUNDAI VELOSTER
JAGUAR XJ
AUDI A8
GOD, YES
ARE YOU A BRAND SNOB?
BENTLEY CONTINENTAL GT CONVERTIBLE
THEY STILL SELL THOSE?
MITSUBISHI LANCER
NAH, WHO CARES
LINCOLN MKZ
INFINITI M37
OLD SCHOOL
KAWAII
VW BEETLE
FIAT 500
SCION iQ
NISSAN CUBE
RUN ON WHAT?
SOME ELECTRICITY
DIESEL
THAT STUFF THAT WORKED SO WELL IN THE HINDENBERG
BATTERIES
CHEVY VOLT
VW JETTA SPORTWAGON TDI
HONDA FCX CLARITY
NISSAN LEAF
FORD FUSION HYBRID
HOW ABOUT A KOREAN TAKE?
SURE, WHY NOT
KIA SOUL
CAN'T A PARENT
OPEN THE HATCH BY KICKING THE AIR
FORD C-MAX
HAVE A COOL CAR
BMW X1
SUBARU WRX
PORSCHE BOXSTER
LOTUS ELISE
PORSCHE CAYMAN
NONE OF YOUR BUSINESS
NO!
HAVE A HAIRPIECE?
NAH
I'M COMFORTABLE
EVER DRIVE ON UNPAVED ROADS?
HOW'S YOUR CASH SITUATION?
SURE DO
AUDI ALLROAD QUATTRO
NOTSO HOTSO
SCION FR-S/SUBARU BRZ
MAZDA MIATA
CATERHAM CLASSIC
WHAT'S FUN FOR YOU?
BEING AWESOME
GOING FAST
CARVING CURVES
GOING SLOW, FEELING FAST
MORGAN 3-WHEELER
HOW MANY CYLINDERS?
WHERE DO YOU WANT YOUR ENGINE?
6
8
12
WANT A STICK?
YEP
ASTON MARTIN VANTAGE
BEHIND ME
IN FRONT OF ME
PORSCHE 911
NISSAN GT-R
HOW MANY PEOPLE ARE WITH YOU?
IS 12 TOO MANY, BUT 8 TOO FEW?
YEP
LAMBORGHINI GALLARDO
SRT VIPER
AUDI R8
I FORGOT, I NEED A USABLEBACK SEAT
0-1
2-3
4+
CADILLAC CTS-V WAGON
WANT TO GO FAST WITH 4+ PEOPLE WITHOUT GAS?
SURE
TESLA MODEL S
BMW 335i
CADILLAC ATS
ARE YOU LOADED?
WANT TO PEEL OUT AT A DAIRY QUEEN?
FUCK YEAH
FORD MUSTANG
CHEVY CAMARO
DODGE CHALLENGER
WHY, YE$
ONLY IF YOU MEAN DRUNK
UM, I'M GOOD
SHOUTY
CLASSY OR SHOUTY?
CHEVROLET CORVETTE STINGRAY
FERRARI 458
McLAREN MP4-12C
CLASSY
JAGUAR F-TYPE
AXIS OR ALLIES?
AXIS
ALLIES
BOTH?
BMW M5
MERCEDES E63 AMG
PORSCHE PANAMERA
CADILLAC CTS-V
MASERATI GHIBLI
FIAT 500 ABARTH
BRING THE NOISE
HOW MUCH DO YOU LIKE NOISE?
SHH, KEEP IT DOWN
MINI COOPER
FORD FIESTA ST
PLUCKY
WHAT "P" EXCITES YOU?
PRECISE
PRACTICAL
PUNCHY
VW GOLF GTI
FORD FOCUS ST
MAZDASPEED3
VOLVO S60
JALOPNIK
CHART BY JASON TORCHINSKY
© 2013

MILESTONE 8 PREFERENCES

- Then we questioned ourselves:
  - ***What about the ordering of facets, terms and objects? Should the user only restrict the focus? Why not allowing the user to change the order based on his/her preferences?***
- Outcome
  - A framework for preferences over multi-dimensional and hierarchical information spaces
  - An extension of the interaction model of faceted search with preferences
  - The Hippalus system that realizes it

System: Hippalus (2013)

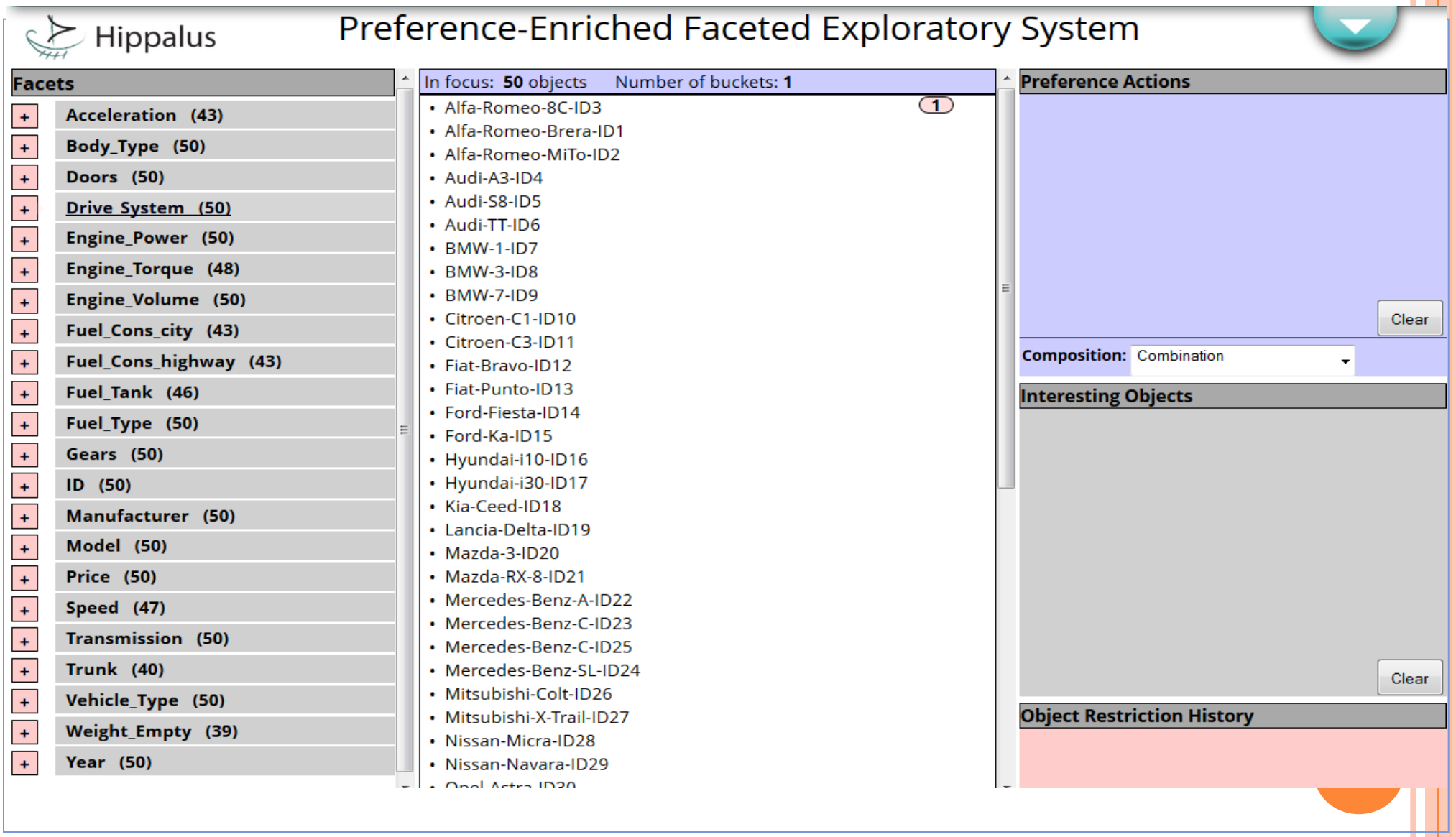
Hippalus: Preference-enriched Faceted Exploration



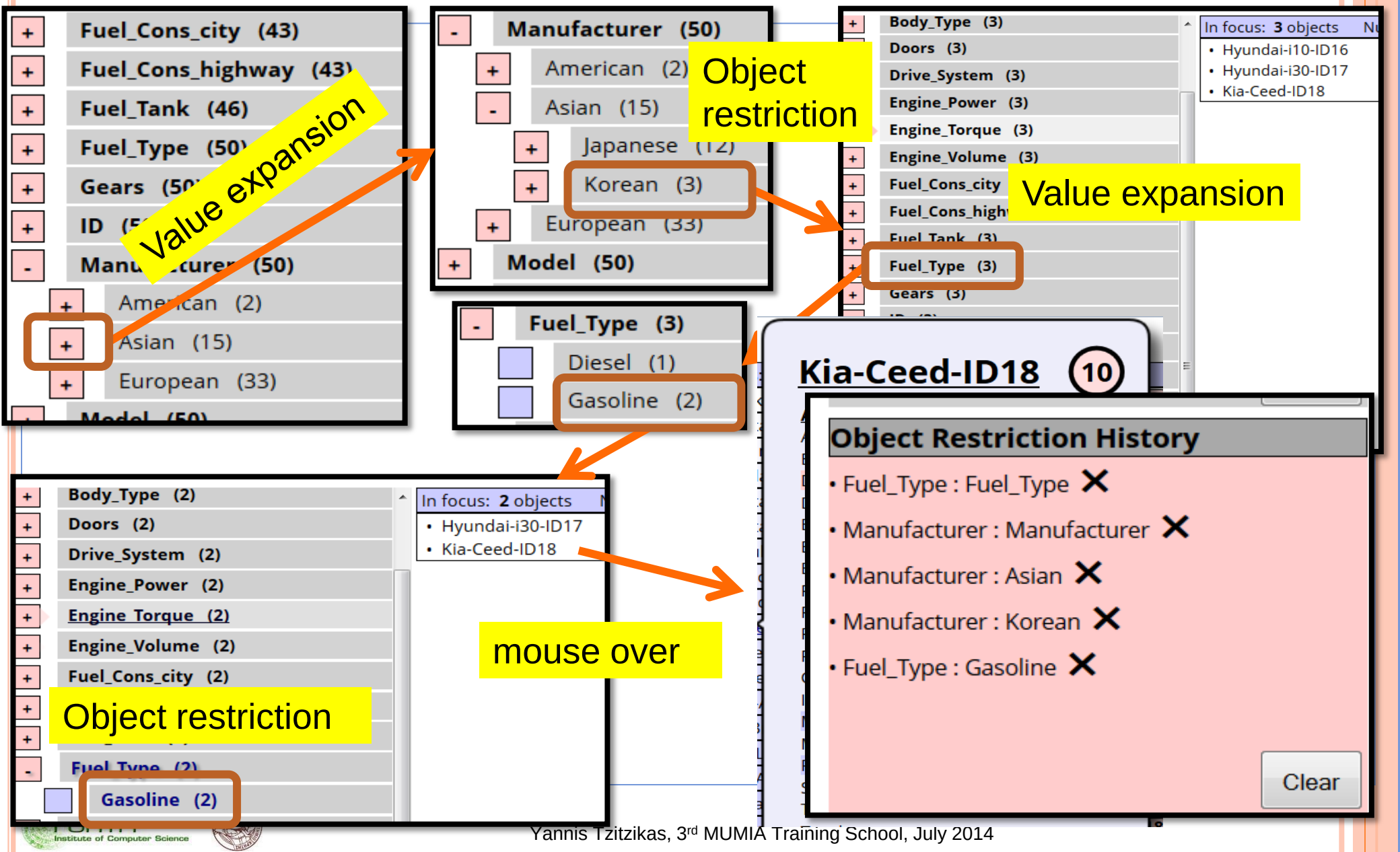
P. Papadakos[1,2] and Y. Tzitzikas[1,2]

- Allows faceted browsing and also supports Preferences
  - User actions specify the ranking of the information space
  - Gradual preference specification
  - Automatic resolution of conflicts
  - Different preference composition modes
    - E.g. if the user defines the desired ordering wrt each dimension, then the first block of the ranked objects is the skyline

HIPPALUS: INTERACTION OVER A KB OF 50 CARS

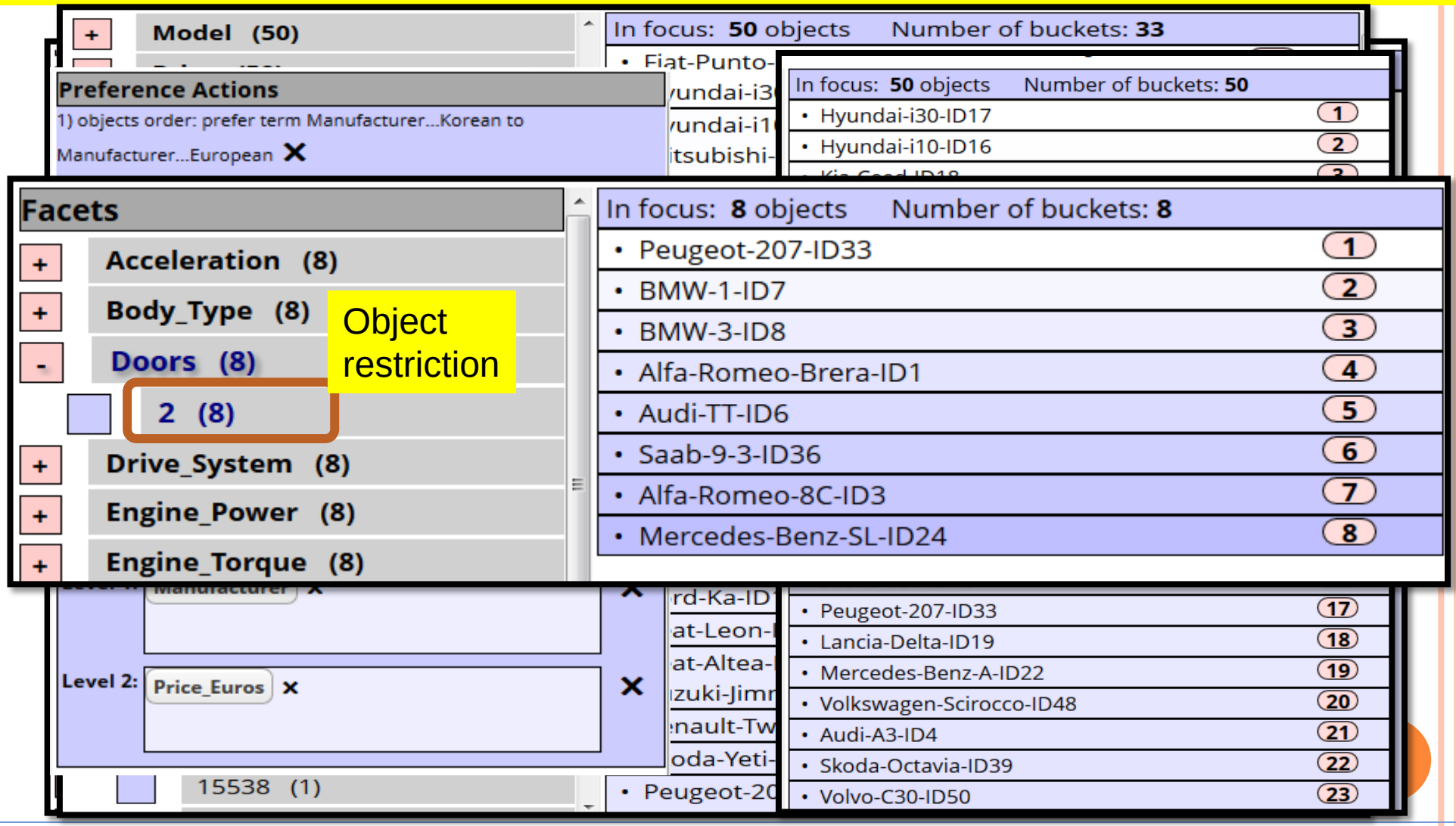


HIPPALUS: FDT INTERACTIONS



HIPPALUS: PREFERENCE ACTIONS

Cars ordered with priority on manufacturer



CONT.

Evaluation

- Over a collection of 50 cars
- With the preference-enriched, all users completed successfully all tasks leading to
  - ideal scores for Precision and Recall!
  - on average in 1/3 of the time!
  - on average with 1/3 of the actions!
- None of the users completed successfully all tasks with the plain interface
- All users (either plain or experts) preferred the preference-enriched interface

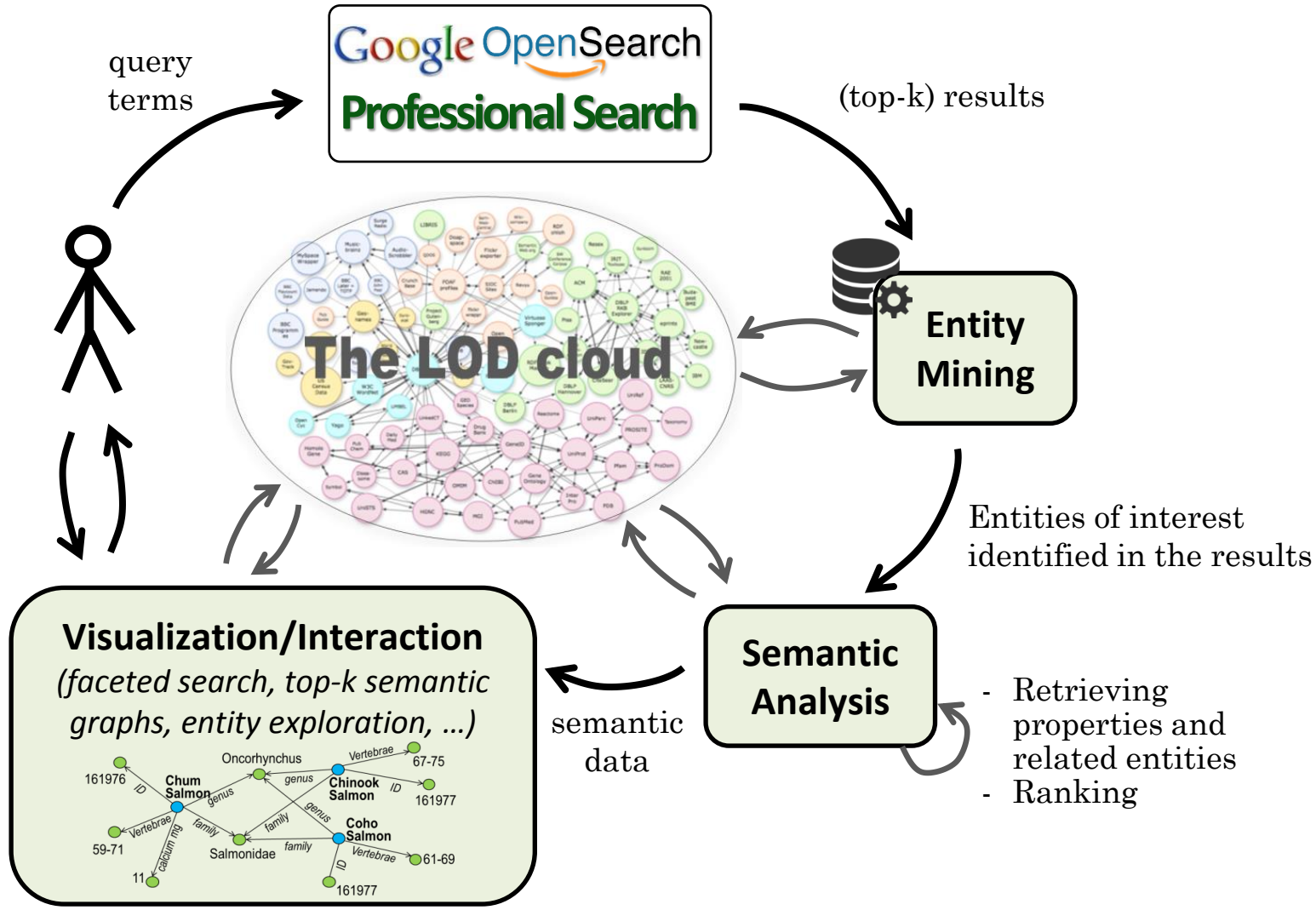
More information in the publications

- [J. FI 13] Yannis Tzitzikas and Panagiotis Papadakos. Interactive Exploration of Multidimensional and Hierarchical Information Spaces with Real-Time Preference Elicitation. In *Journal FUNDAMENTA INFORMATICAE*, 2013
- [ExploreDB'14] Panagiotis Papadakos, Yannis Tzitzikas: Hippalus: Preference-enriched Faceted Exploration. In EDBT/ICDT Workshops 2014

MILESTONE 9 FROM DIMENSIONS TO GRAPHS

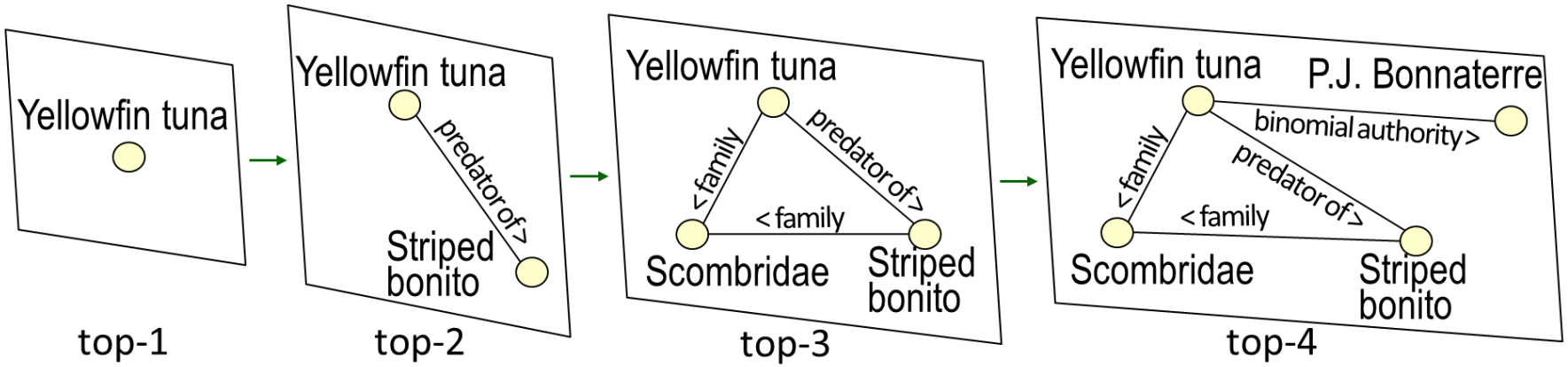
- Then we questioned ourselves:
  - ***So far we have seen services for getting and exploiting multidimensional spaces over the search results. But what if the notion of dimension cannot be defined, or in case there are too many? What can be done without having to configure entity types?***
- Outcome
  - A semantic post-processing of results that does not yield a multidimensional space but a graph.
- Challenges
  - Graph construction and exploitation for identifying the important (useful for the user) nodes and relationships

TOP-K SEMANTIC GRAPHS



TOP-K SEMANTIC GRAPHS (CONT.)

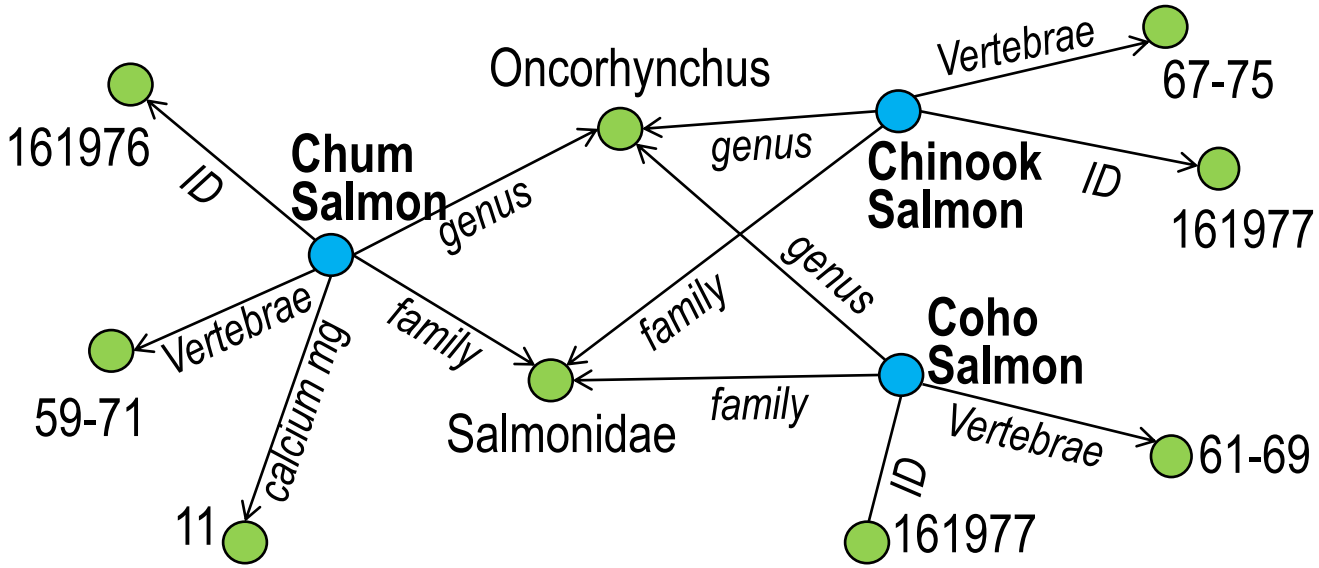
- The system can return the top-K graph for any K from 1 to number of nodes produced
  - Vertices: the K most highly ranked nodes
  - Edges: the edges that connect the K most highly ranked nodes
- The user is free to increase or reduce the value of K
- Example (from a real domain):



TOP-K SEMANTIC GRAPH

This graph

- can **complement** the query answer with useful information regarding the connectivity of the identified entities
- allows users to **instantly inspect** information that may lie in different places and that may be laborious and time consuming to locate
- provides useful information about the **context** of the identified entities
- allows the users to get a **more sophisticated overview** and to make better sense of the results



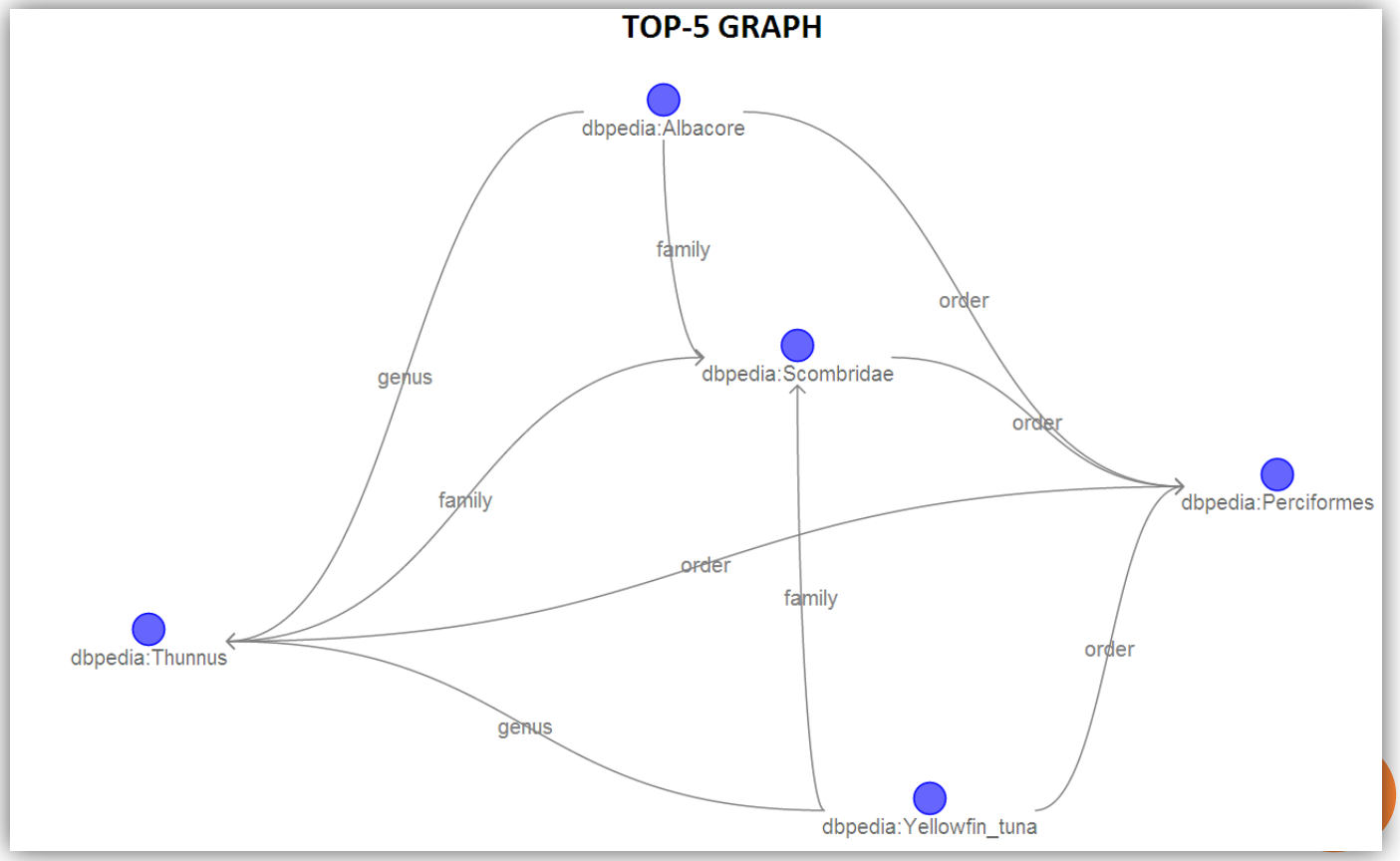
PROTOTYPE (2014)

http://139.91.183.72/x-ens-2/

yellowfin tuna | Search

TOP-5 LIST

1. dbpedia:Yellowfin_tuna
2. dbpedia:Perciformes
3. dbpedia:Scombridae
4. dbpedia:Albacore
5. dbpedia:Thunnus



CONT.

- Evaluation (main results)
  - Usefulness – Survey for the marine domain
    - *The majority of participants believe that the appearance of a graph of semantic information related to the search results can help them during an exploratory search process*
  - Effectiveness – Comparative evaluation of ranking schemes:
    - The proposed PageRank-based ranking scheme produces more preferred ranking compared to other link analysis-based algorithms
  - Efficiency – Case study over online DBpedia
    - The exploitation of LOD can be supported at query-time
    - For up to 100 detected entities we can offer the proposed functionality at real-time, even if we query an online KB (like DBpedia)
  - The major bottleneck is the reliability and performance of online SPARQL endpoints
    - We expect this limitation to get overcome in the near future
    - In the meanwhile, we can use caching / indexing / dedicated warehouses / distributed infrastructure
- Related Publications:
  - P. Fafalios and Y. Tzitzikas, Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time, IEEE 8th International Conference on Semantic Computing (ICSC'14), Newport Beach, California, USA, June 2014

5. Synopsis and Discussion (30')

Synopsis

Challenges

Synopsis

- We have discussed information needs of exploratory nature
- We have seen the basics of faceted exploration
- We have seen an overview of semantic technologies
- We have seen ways to exploit semantic datasets during searching (exploratory search) with emphasis on doing this at search time

- So what's next ?

Directions & Challenges

- **Ubiquity.**
  - Faceted browsing of search results and gradual restriction should be possible for any kind of query, for any domain and with no predetermined facets.
  - In other words, methods that bypass the need for explicit configuration (regarding facets, entities types, LOD sources) are required.
    - This is why we currently study M9
      - Then we also have to define the interaction model over such graphs

Directions & Challenges (2)

- **Fusion of Structured and Unstructured Content.**
  - The exploitation of LOD in the exploratory search process is promising, e.g. for Named Entity Recognition and disambiguation. However, the fusion of structured and unstructured content requires more work.
    - E.g. should (and if yes how) the results of the semantic post processing should affect the search hits (a kind of semantic feedback)

Directions & Challenges (3)

- **User Control.**
  - Explicit, user-provided and controllable preference management is beneficial for supporting a transparent decision making process.
  - We believe that the framework supported by the Hippalus system is a first step towards this direction.

Directions & Challenges (4)

- **Evaluation.**
  - We need easy to follow methods for evaluating the effectiveness of exploratory search methods, and easily reproducible evaluation results.
  - Although the classical IR has well established methodologies for evaluation, things are not so clear and straightforward in interactive IR (IIR).

Questions and Exercises

References and Links

Organized on the basis of the case studies

References and Links

- Faceted Search and Dynamic Taxonomies
  - Sacco, Giovanni Maria; Tzitzikas, Yannis (Eds.), Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience, Series: The Information Retrieval Series , Vol. 25, 2009
  - Giovanni Maria Sacco: **Dynamic Taxonomies: A Model for Large Information Bases.** IEEE Trans. Knowl. Data Eng. 12(3): 468-479 (2000)
- Browsing Approach for (plain and fuzzy) RDF
  - Sébastien Ferré's Publications
  - Nikos Manolis, Yannis Tzitzikas: **Interactive Exploration of Fuzzy RDF Knowledge Bases.** ESWC (1) 2011: 1-16

REFERENCES (M1)

- **Mitos**-related (Web Searching and Faceted Search)
  - [PCI'08] Panagiotis Papadakos, Yannis Theoharis, Yannis Marketakis, Nikos Armenatzoglou and Yannis Tzitzikas.Mitos: Design and Evaluation of a DBMS–based Web Search Engine. In PCI 2008.
  - [FIND'08] Yannis Tzitzikas, Nikos Armenatzoglou, and Panagiotis Papadakos. FleXplorer: A Framework for Providing Faceted and Dynamic Taxonomy-based Information Exploration. In DEXA/FIND 2008.
  - [ECDL'09] Panagiotis Papadakos, Stella Kopidaki, Nikos Armenatzoglou and Yannis Tzitzikas. Exploratory Web Searching with Dynamic Taxonomies and Results Clustering. In ECDL 2009.
  - [WISE'09] Stella Kopidaki, Panagiotis Papadakos, and Yannis Tzitzikas. STC+ and NM-STC: Two novel online results clustering methods for web searching. In WISE 2009.
  - [J. KAIS 2012] Panagiotis Papadakos, Stella Kopidaki, Nikos Armenatzoglou and Yannis Tzitzikas. On exploiting Static and Dynamically mined Metadata for Exploratory Web Searching. In KAIS Journal 2012.

REFERENCES (M2)

- Instance Overview Search-related
  - Pavlos Fafalios, Yannis Tzitzikas: Exploiting Available Memory and Disk for Scalable Instant Overview Search. WISE 2011: 101-115
  - P. Fafalios, I. Kitsos and Y. Tzitzikas, Scalable, Flexible and Generic Instant Overview Search, Proceedings of the 21st International Conference on World Wide Web (demo paper), WWW 2012, Lyon, France, April 2012
- Links to Online Prototypes
  - http://www.ics.forth.gr/ios

REFERENCES (M3+M4)

- Semantic-Post Processing of Search Results with Entity Mining and LOD
  - P. Fafalios, I. Kitsos, Y. Marketakis, C. Baldassarre, M. Salampasis and Y. Tzitzikas, Web Searching with Entity Mining at Query Time, Proceedings of the 5th Information Retrieval Facility Conference, IRF 2012, Vienna, July 2012
  - Pavlos Fafalios, Yannis Tzitzikas: X-ENS: semantic enrichment of web search results at real-time. SIGIR 2013, Dublin, Ireland, 28 July - 1 August 2013
- Links to Online Prototypes
  - http://62.217.127.118/x-ens/

REFERENCES (M5)

- Semantic-Post Processing of Search Results with Entity Mining in **Patent Search**
  - P. Fafalios and Y. Tzitzikas, "Exploratory Professional Search through Semantic Post-Analysis of Search Results", In "Professional Search in the Modern World", Lecture Notes in Computer Science (LNCS), Springer, 2014 (accepted for publication as a State-of-the-Art volume in LNCS)
  - P. Fafalios, M. Salampasis and Y. Tzitzikas, *Exploratory Patent Search with Faceted Search and Configurable Entity Mining,* 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with ECIR'13, Moscow, Russia, March 2013
- Links to Online Prototypes
  - http://www.perfedpat.eu/index.php/download-perfedpat
  - http://139.91.183.72/x-search-metadata-groupings/

REFERENCES (M6)

- Semantic-Post Processing of Search Results with Entity Mining in **Marine Search**
  - P. Fafalios and Y. Tzitzikas, "Exploratory Professional Search through Semantic Post-Analysis of Search Results", In "Professional Search in the Modern World", Lecture Notes in Computer Science (LNCS), Springer, 2014 *(accepted for publication as a State-of-the-Art volume in LNCS)*.
- Links to Online Prototypes
  - http://62.217.127.118/x-search/
  - http://62.217.127.118/x-search-fao/

REFERENCES (M7)

- Semantic-Post Processing of Search Results with Entity Mining in **over the Cloud**
  - I. Kitsos, K. Magoutis and Y. Tzitzikas, Scalable Entity-based Summarization of Web Search Results using MapReduce, Journal on Distributed and Parallel Databases (DAPD), 2014

REFERENCES (M8)

- Extending Faceted Search with **Preferences**
  - [WISE'12] Panagiotis Papadakos, Yannis Tzitzikas and Dafni Zafeiri: An Interactive Exploratory System with Real-Time Preference Elicitation. Demo paper, In *Web Information Systems Engineering - WISE 2012*, Pafos, Cyprus, Volume 7651, p. 817-820, November 2012.
  - [J. FI 13] Yannis Tzitzikas and Panagiotis Papadakos. Interactive Exploration of Multidimensional and Hierarchical Information Spaces with Real-Time Preference Elicitation. In *Journal FUNDAMENTA INFORMATICAE*, Volume 122, Issue 4, pp 357-399, 2013.
  - [ExploreDB'14] Panagiotis Papadakos, Yannis Tzitzikas: Hippalus: Preference-enriched Faceted Exploration. EDBT/ICDT Workshops 2014: 167-172
- Video Demonstration available at http://www.youtube.com/watch?v=Cah-z7KmlXc
- Links to Online Prototypes http://www.ics.forth.gr/isl/Hippalus

REFERENCES (M9)

- Graph-resulting Semantic Post-Processing of Search
  - P. Fafalios and Y. Tzitzikas, Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time, IEEE 8th International Conference on Semantic Computing (ICSC'14), Newport Beach, California, USA, June 2014

- Prototype (configured for the marine domain)
  - http://139.91.183.72/x-ens-2/

ACKNOWLEDGEMENTS

- Apart from Panagiotis Papadakos and Pavlos Fafalios, other students have also contributed to the "story" that was presented:
  - Nikos Armenatzoglou
  - Stella Kopidani
  - Nikos Manolis
- These slides include material from
  - The LOD integration scenario was taken from
    - http://www.csee.umbc.edu/courses/graduate/691/spring14/01/notes/01_introduction/01motivating_example.pptx

Thanks for you attention