Exploiting Linked Data in Exploratory Search

Pavlos Fafalios

Thesis Defense

Foundation for Research and Technology – Hellas (FORTH)
Institute of Computer Science (ICS)
Information Systems Laboratory (ISL)

University of Crete
Computer Science Department
Greece

Outline

- Introduction
- Context
- Background and Related Work

≈ 10 min

- The Approach
  - **(A) Configuring Named Entity Extraction**
  - **(B) Stochastic Ranking of Entities and Properties**
  - **(C) Stochastic Re-Ranking of Search results**
- Applications

≈ 30 min

- Conclusion
  - Synopsis of Contributions
  - Directions for Future Research

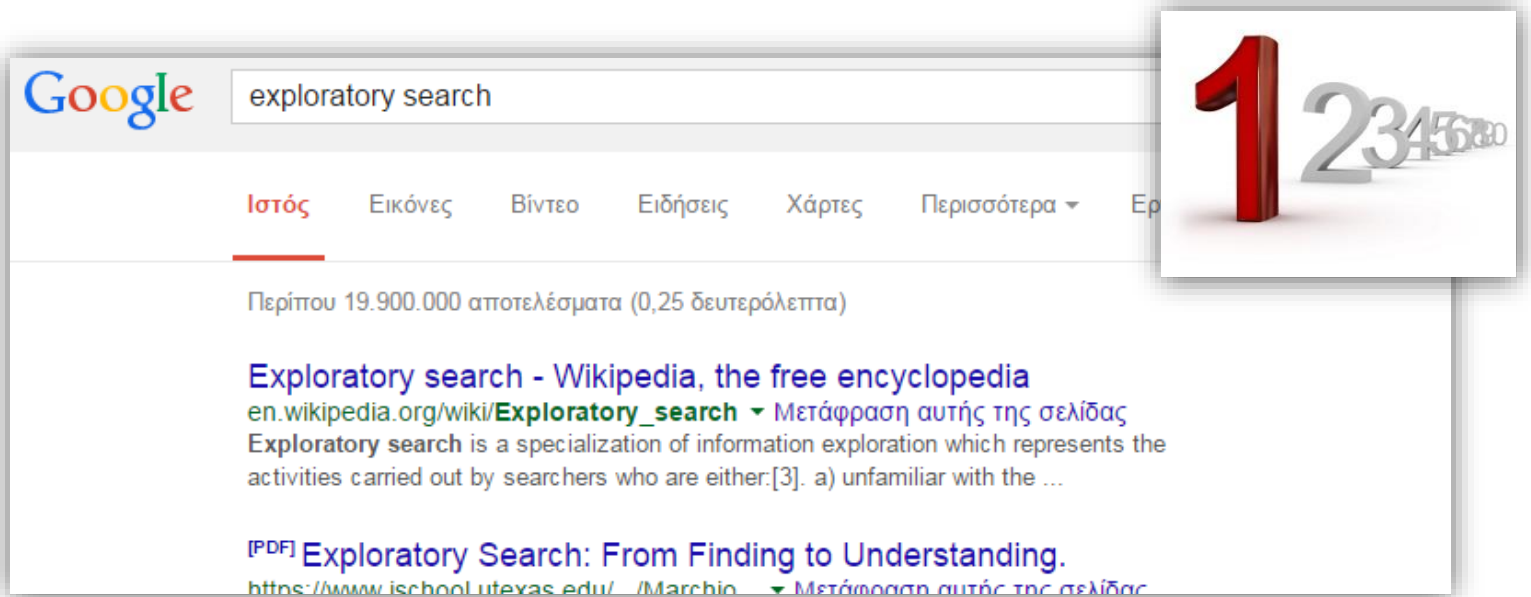
≈ 5 min

Introduction

- Motivation and Vision
- Research Questions
- Contributions

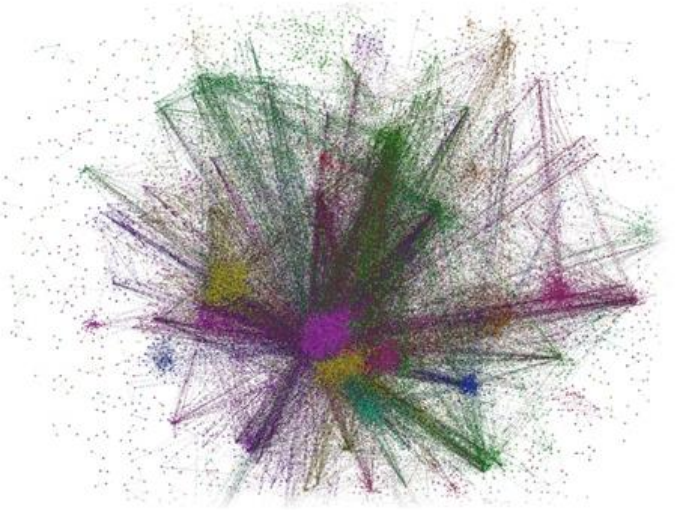
Searching for Information

- Integral part of our daily lives!
  - Google processes over 40,000 search queries every second!
- Most search methods are appropriate for focalized search
  - Users know exactly what they are looking for
- However, a high percentage of search tasks are **exploratory**
  - The canon in **Professional Search**!
  - Common IR techniques are often not enough!

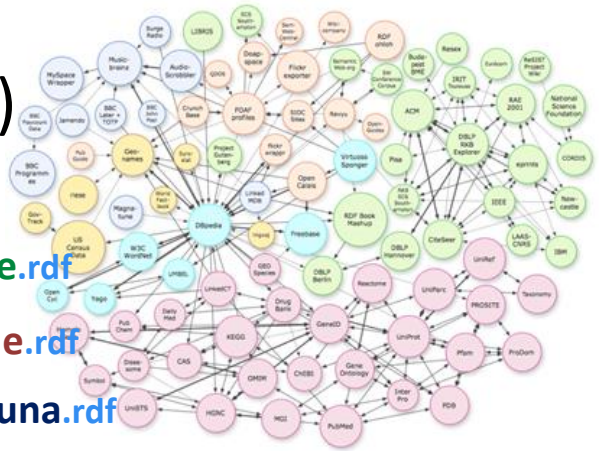


The two “Worlds”

- Web of Unstructured Documents
  - Web pages, online PDF/Text files, …



- Web of Structured Data (Linked Open Data)
  - XML/RDF



http://dbpedia.org/data/**Crete**.rdf

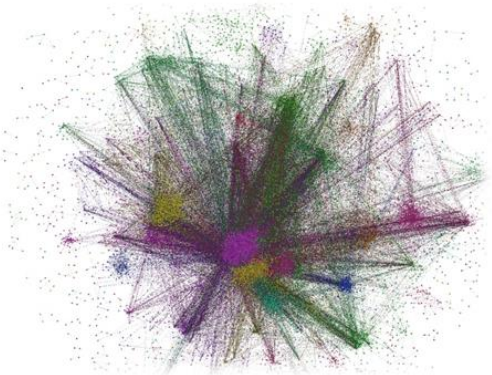
http://dbpedia.org/data/**Aristotle**.rdf

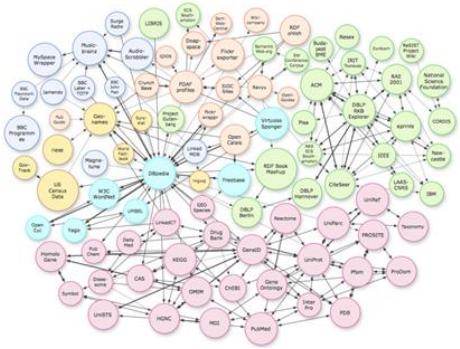
http://dbpedia.org/data/**Tuna**.rdf

http://dbpedia.org/data/**Paracetamol**.rdf

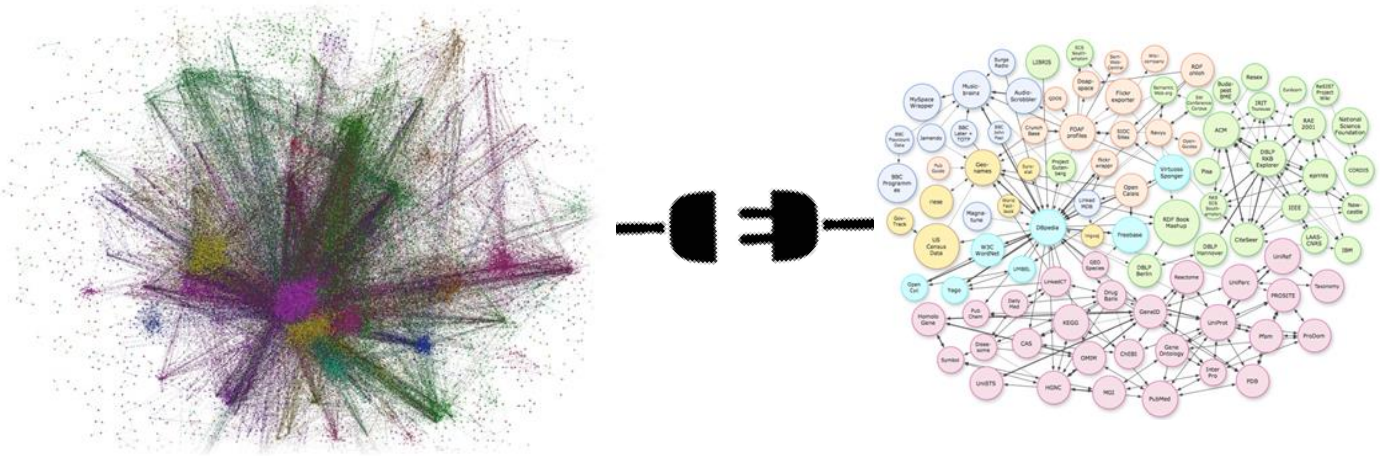
Named-entities

- Elements in text belonging to pre-defined categories
  - names of **Persons**, **Locations**, …, **Actors**, **Countries**, **Cities**, …
  - names of **Species**, **Water Areas**, …, **Diseases**, **Drugs**, **Proteins**, …
  - …
- Occurring in both worlds, in many different kinds of artifacts
  - Web pages
  - Documents
  - Databases
  - RDF statements
  - …





Connecting the two Worlds
(during **information seeking**)

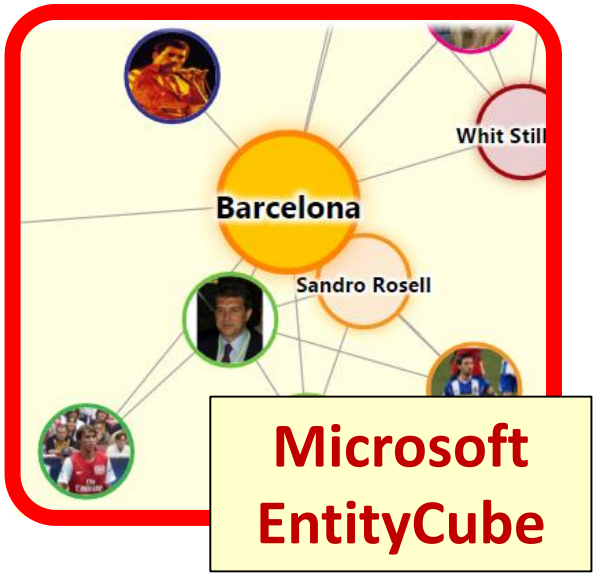


Entity-based integration:
***named entities** are used as the "glue"*
for automatically connecting
documents (search results)** with **data and knowledge

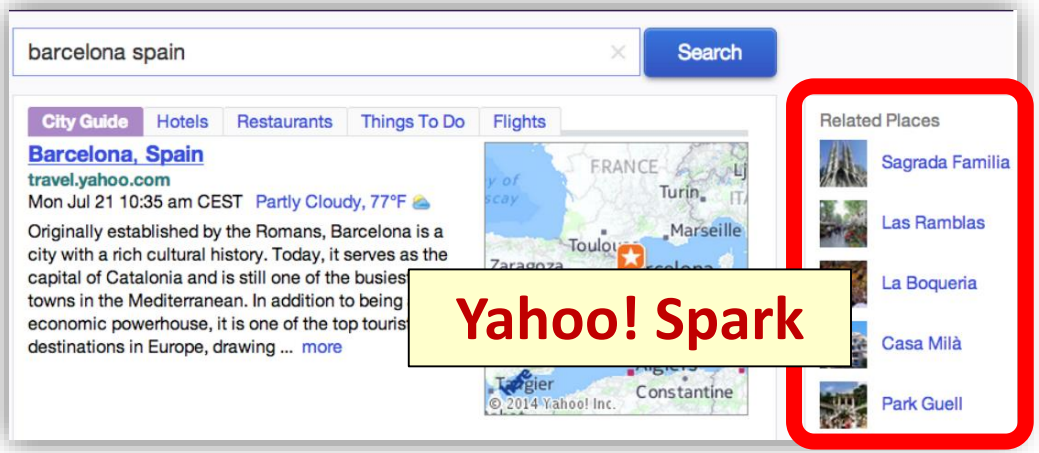
Connecting the two Worlds
(during **information seeking**)

Entity Characteristics

Entity Relationships

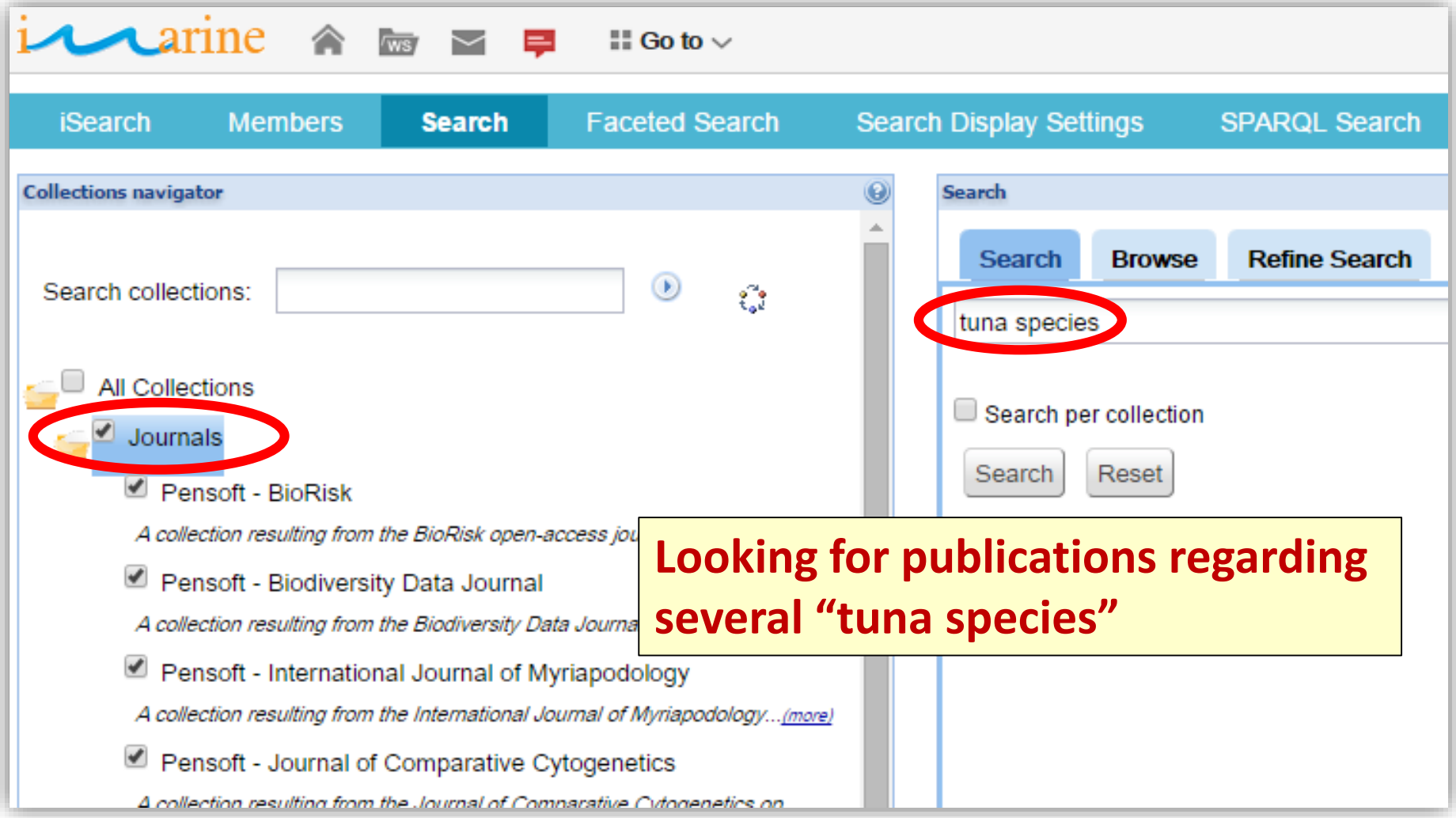


Entity Recommendations

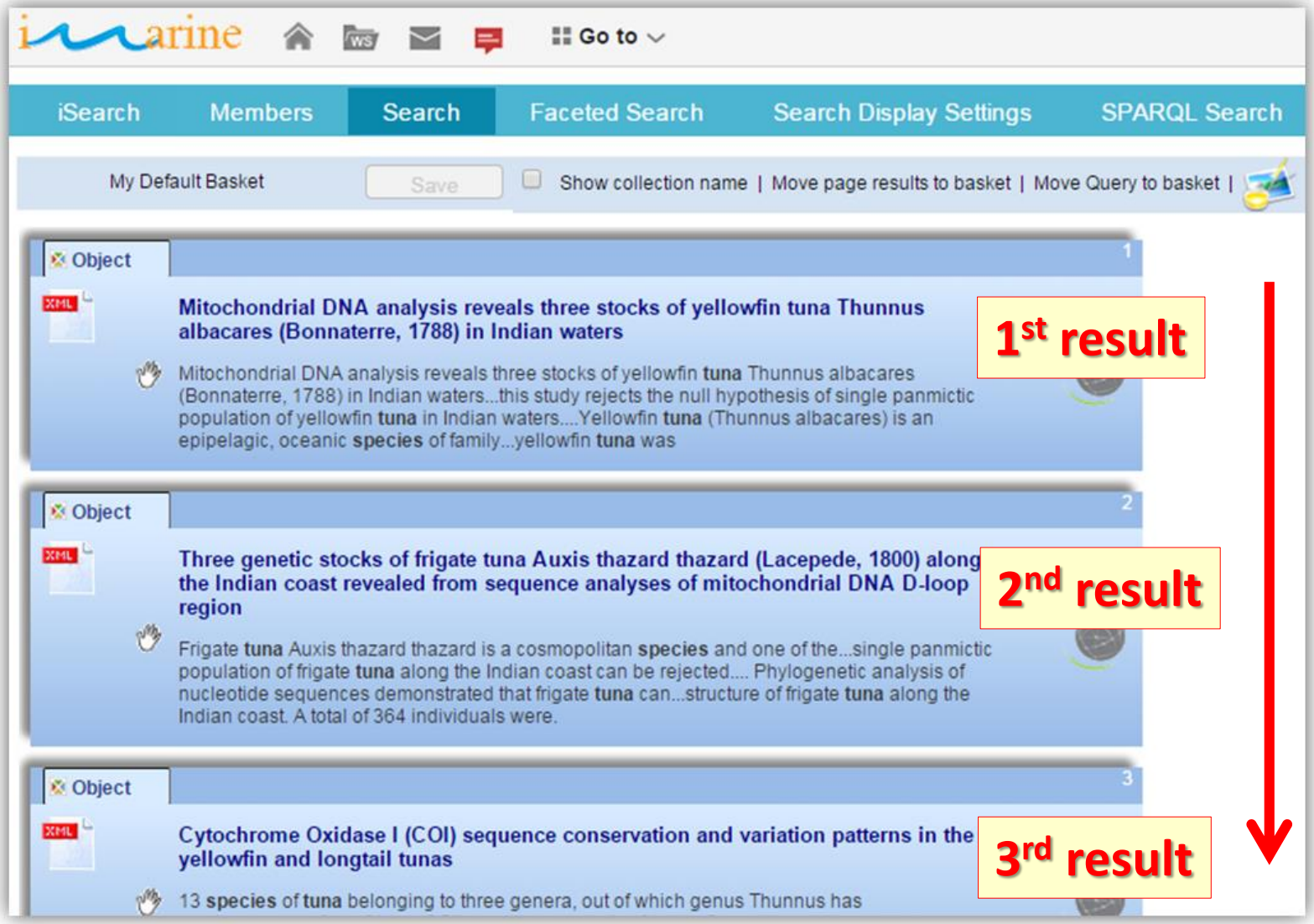


FORTH Institute of Computer Science

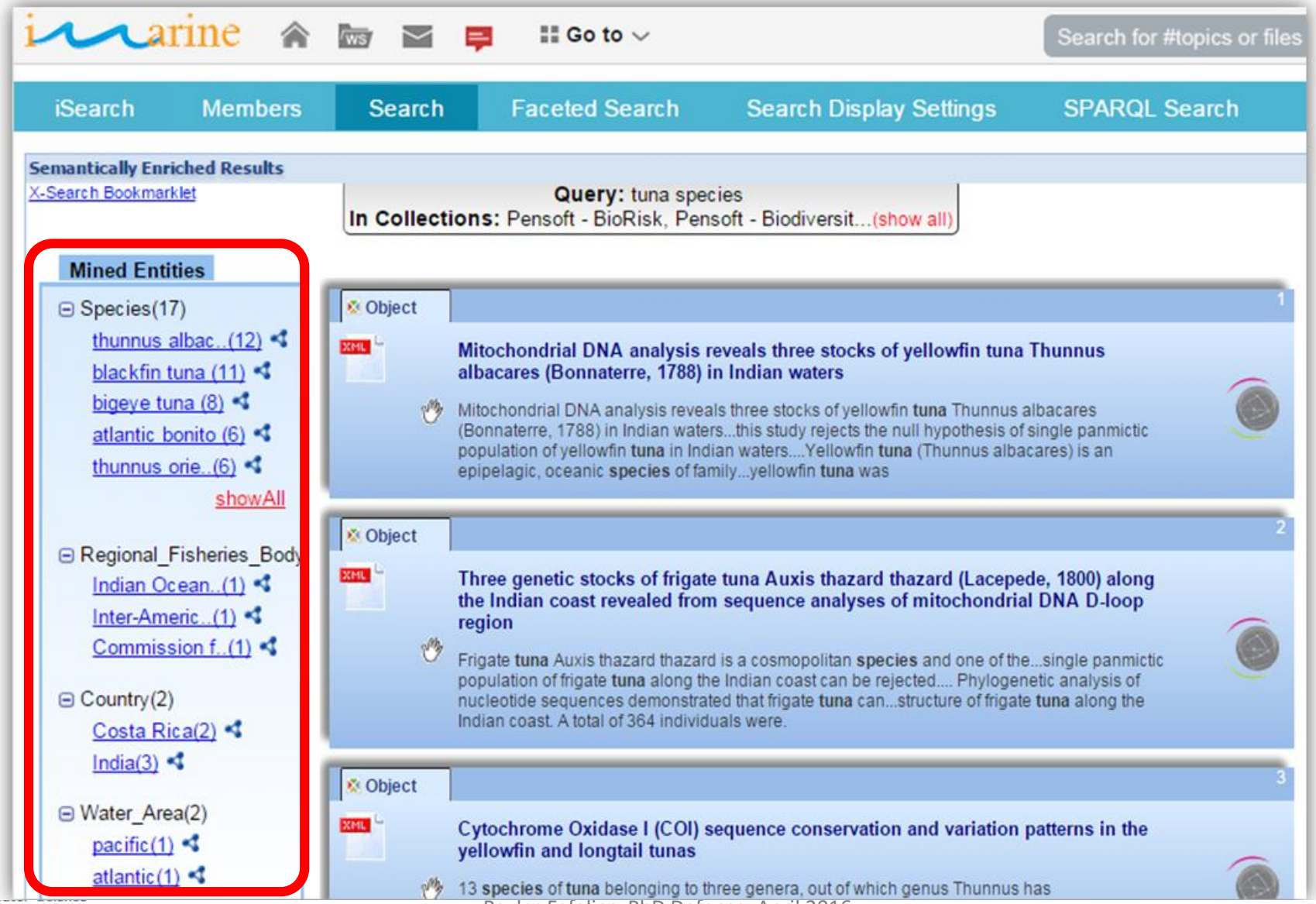
Motivating Scenario – Marine Search



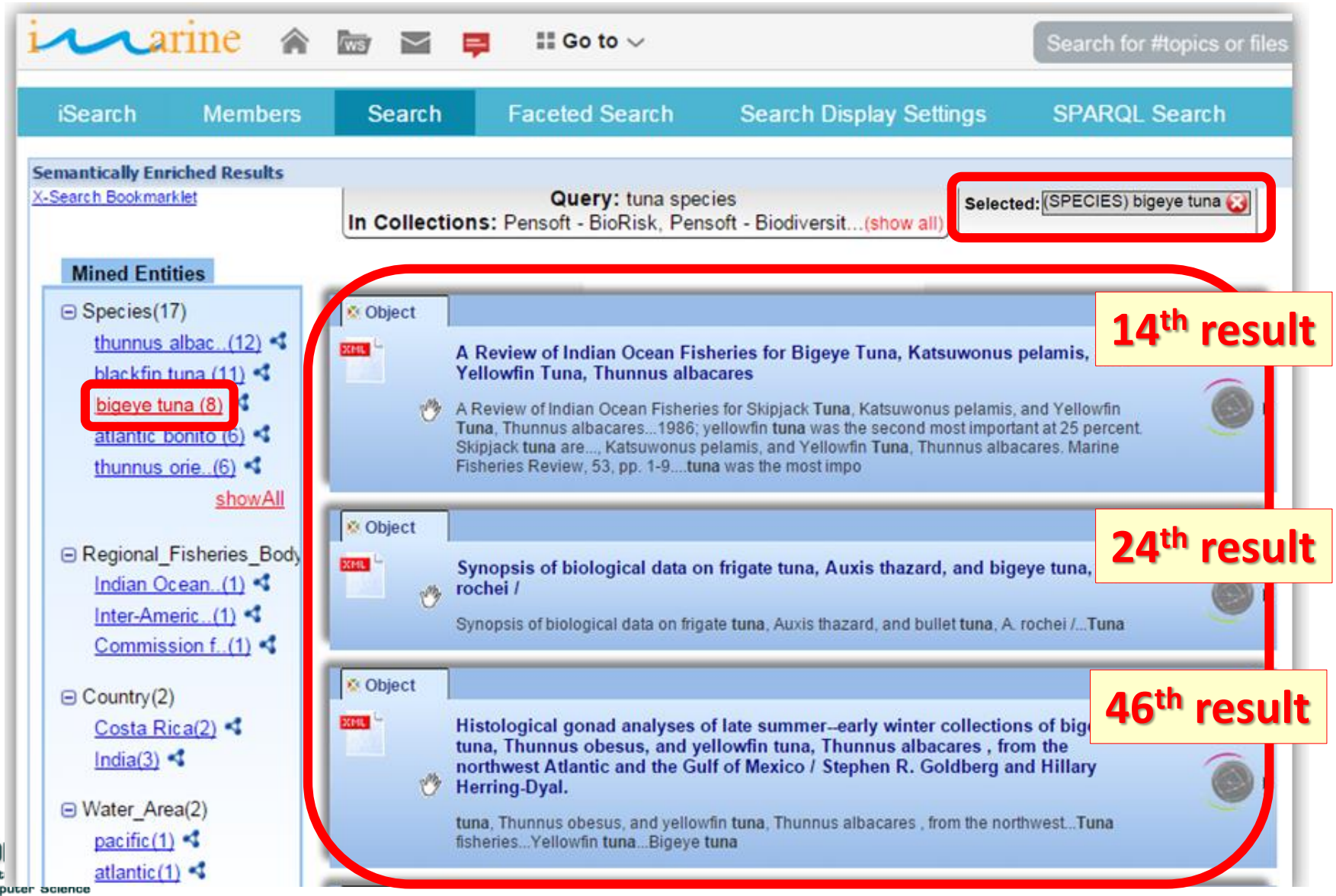
Motivating Scenario – Marine Search



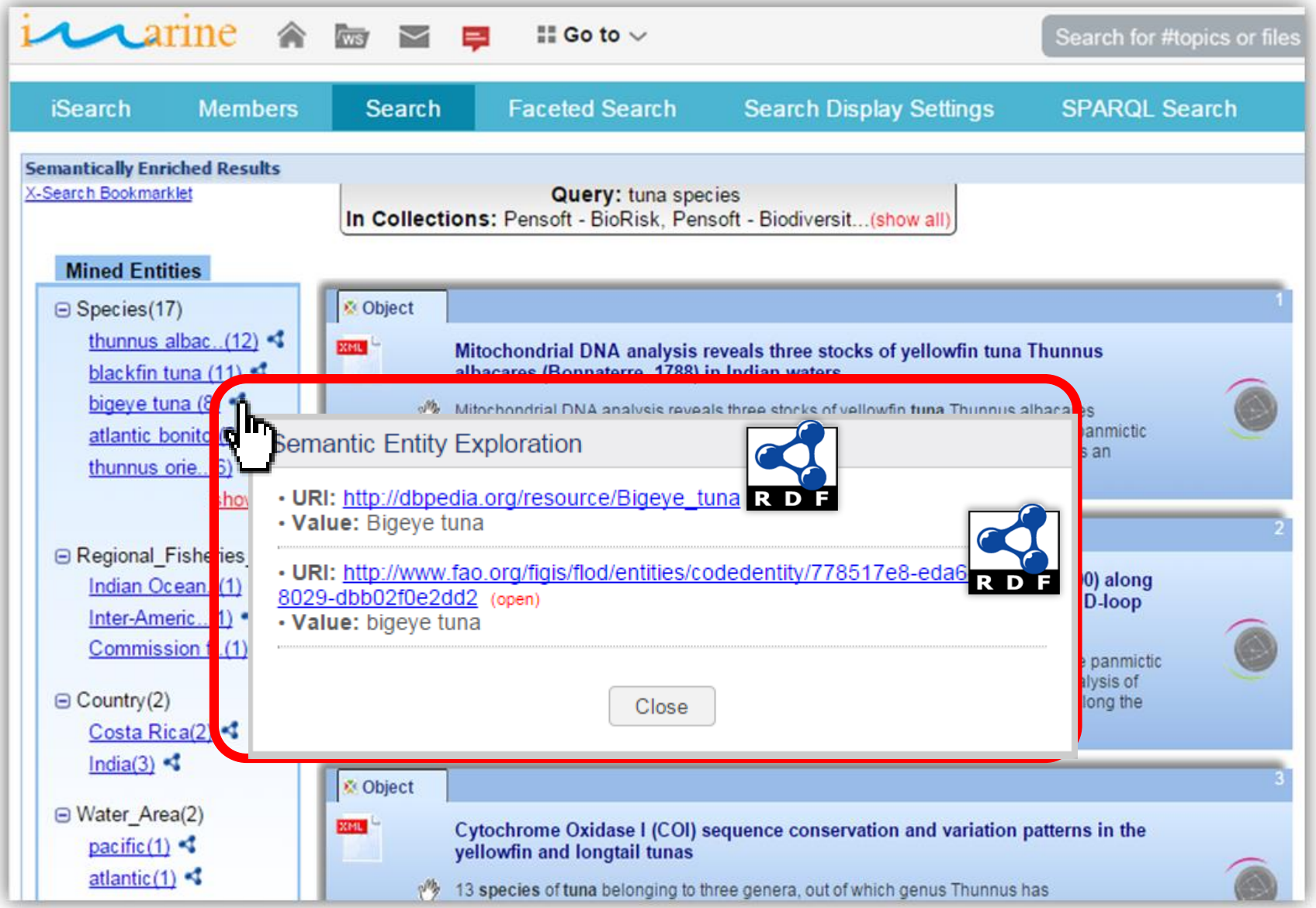
Motivating Scenario – Marine Search



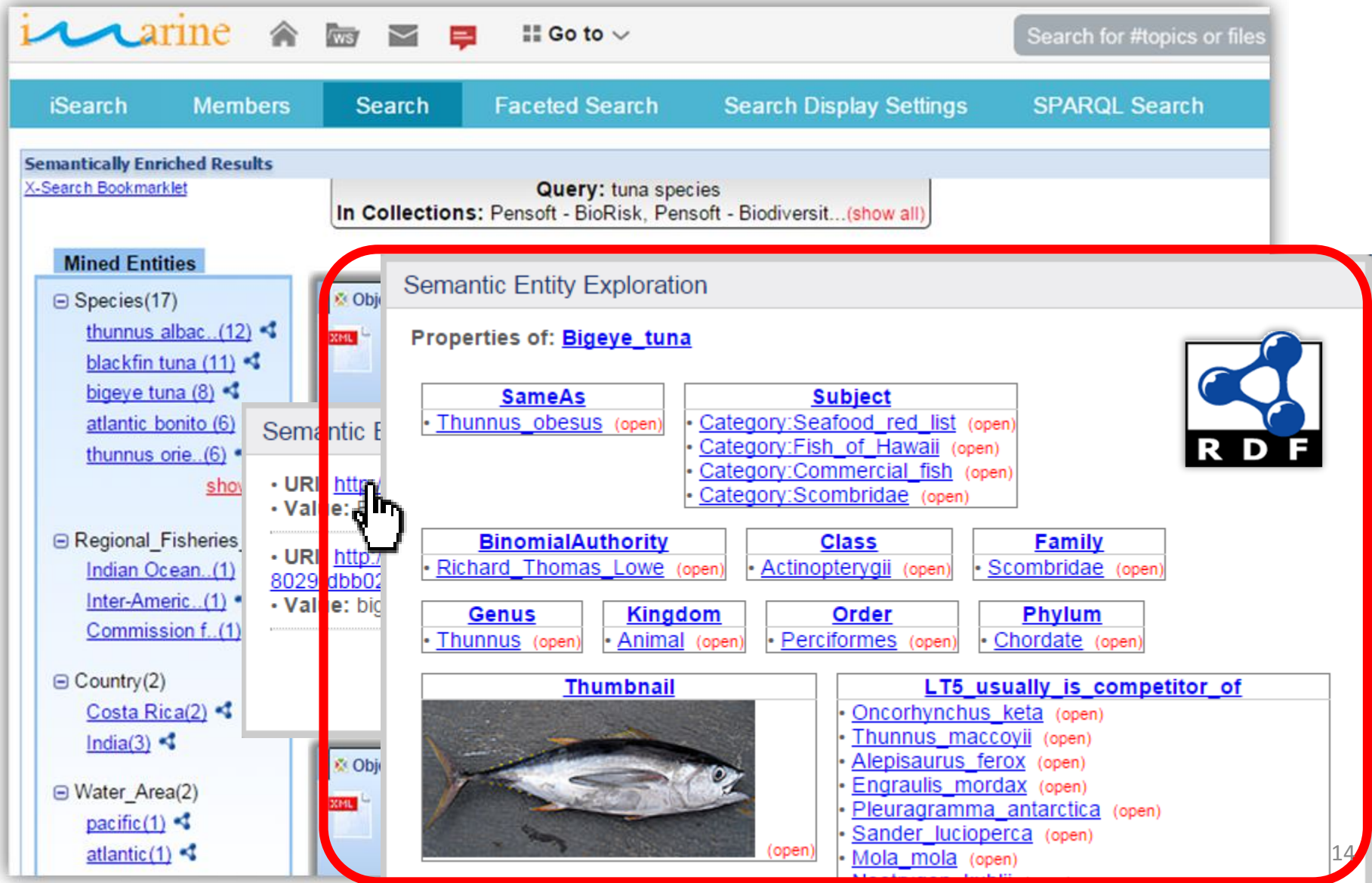
Motivating Scenario – Marine Search



Motivating Scenario – Marine Search



Motivating Scenario – Marine Search



The Problems

- The number of identifiable (extractable) entities can be very high
- The amount of structured information that is available for these entities can be very high too

❖ Need of **specifying** the "**data of interest**"
  - i.e. the *Entities of Interest*, the *Knowledge Bases* containing data about these entities, the *Semantic Information* about the entities that is useful to provide

- The number of entities identified in the search results can be high
- The amount of semantic information that has been retrieved for these entities can be very high too (their associations and properties)

❖ Need of **ranking** the derived semantic information
  - **Identify** and **present** to the user the most important (to the search context) entities, associations and properties

- Ranking is always important
- Classic IR techniques often do not work well for more complex information needs

❖ Need of **exploiting** one or more aspects of the **user** and/or **search context**
  - **Offer** add-on services that can improve the search results

Research Questions

Semantic Context Specification **A**

- How one can **specify** the "entities of interest" (for a given domain)?
- How to specify the additional semantic information about these entities that is useful to provide in a search context?
- Can we dynamically configure all the above and then fetch and exploit such semantic information at query-time?

Semantic Context Selection **B**

- How to **select** the entities (as well as entity associations and properties) that better characterize the search results?
- Can this analysis and selection be done at query-time?

Semantic Context Exploitation **C**

- Can we **exploit** the extracted entities (and the related semantic information) also in pure IR (e.g., for producing a better ranking for the search results)?

Contributions

(A) Configuring Named Entity Extraction

A

Semantic Context Specification

- We propose a generic (abstract) **model** that exploits the **Linked Open Data** (**LOD**) for configuring a **Named Entity Extraction** (**NEE**) service
  - Genericity / Interoperability
- We introduce an RDF/S vocabulary, called "**Open NEE Configuration Model**"
  - Allows a NEE service to describe and publish its entity mining capabilities
- We propose an **extension** of the **Open Annotation Data Model**[1]
  - Allows relating the output of a NEE process with an applied configuration
- We present and evaluate **X-Link**, a NEE framework that adopts the proposed configuration model

[1] W3C Community Draft, 2013 (http://www.openannotation.org)

Contributions

(B) Stochastic Ranking of Entities and Properties

Semantic Context Selection

B

- We introduce a **probabilistic** scheme for **ranking** entities and properties
  - Identifies the "**important**" semantic data that better characterizes the search results
  - Produces **Top-K Semantic Graphs** related to the search results
  - Comparative evaluation with other ranking methods
  - Experimental Results on Efficiency

Contributions

(C) Stochastic Re-Ranking of Search Results

Semantic Context Exploitation

C

- We introduce a biased **Random Walk** model that exploits the semantic context for **re-ranking** the list of retrieved results
  - Promotes low-ranked but relevant hits containing important (for the search context) entities
  - Evaluation using **TREC** (**T**ext **RE**trieval **C**onference) datasets

Context

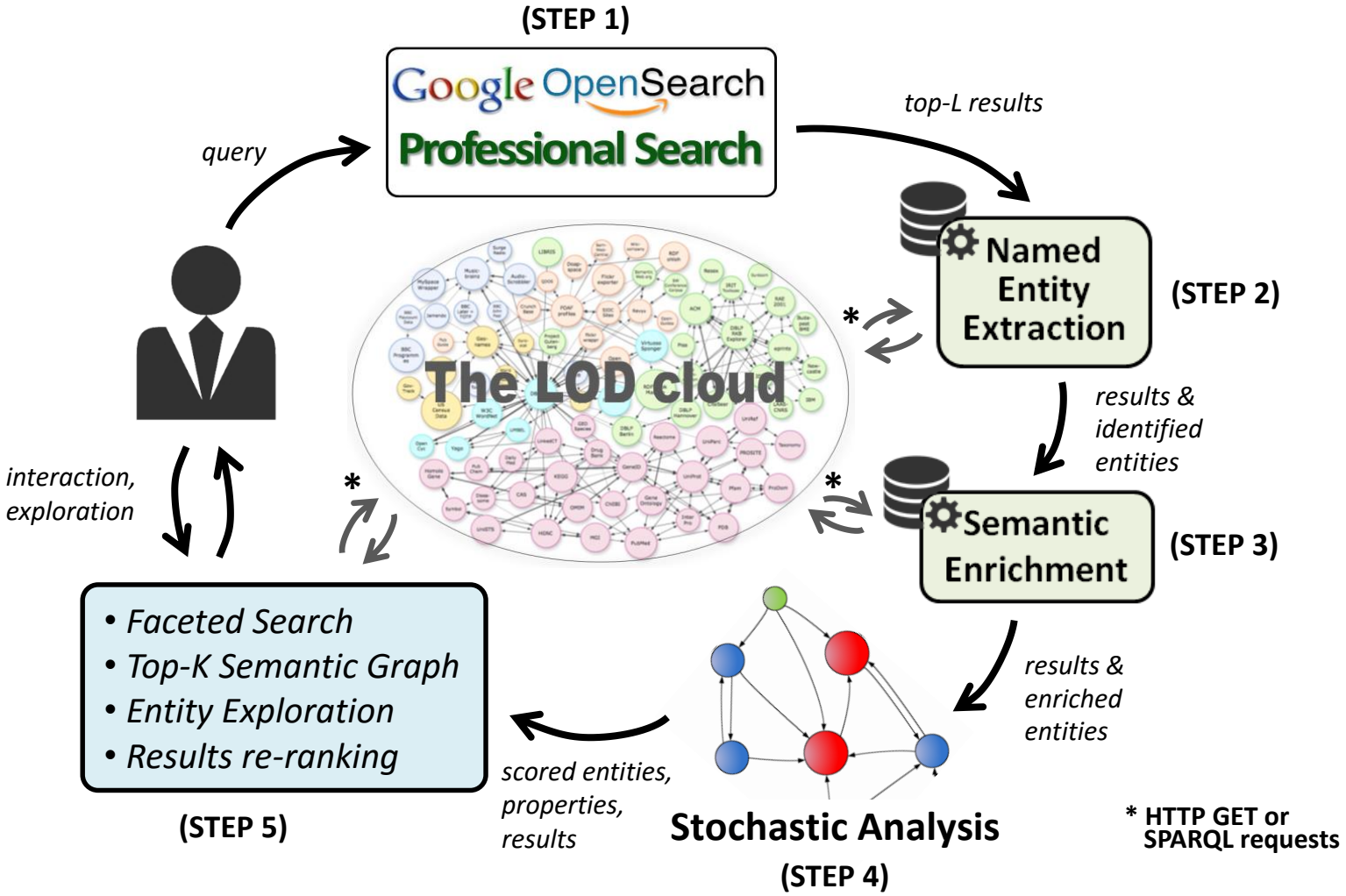
- Main Notions
- The considered Search Process
- The considered Configuration Process

Main Notions

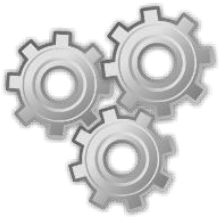
- **Entities of Interest (EoI)**
  - Named entities that are meaningful in the application context
  - E.g.:
    - For a **marine**-related application: *fish species*, *countries*, *water areas*,…
    - For a **medicine**-related application: *drugs*, *diseases*, *proteins*, …
- **Semantic Knowledge Base (SKB)**
  - Semantic repository of RDF data accessible as LOD or through a SPARQL endpoint
  - DBpedia, DrugBank, YAGO2, …

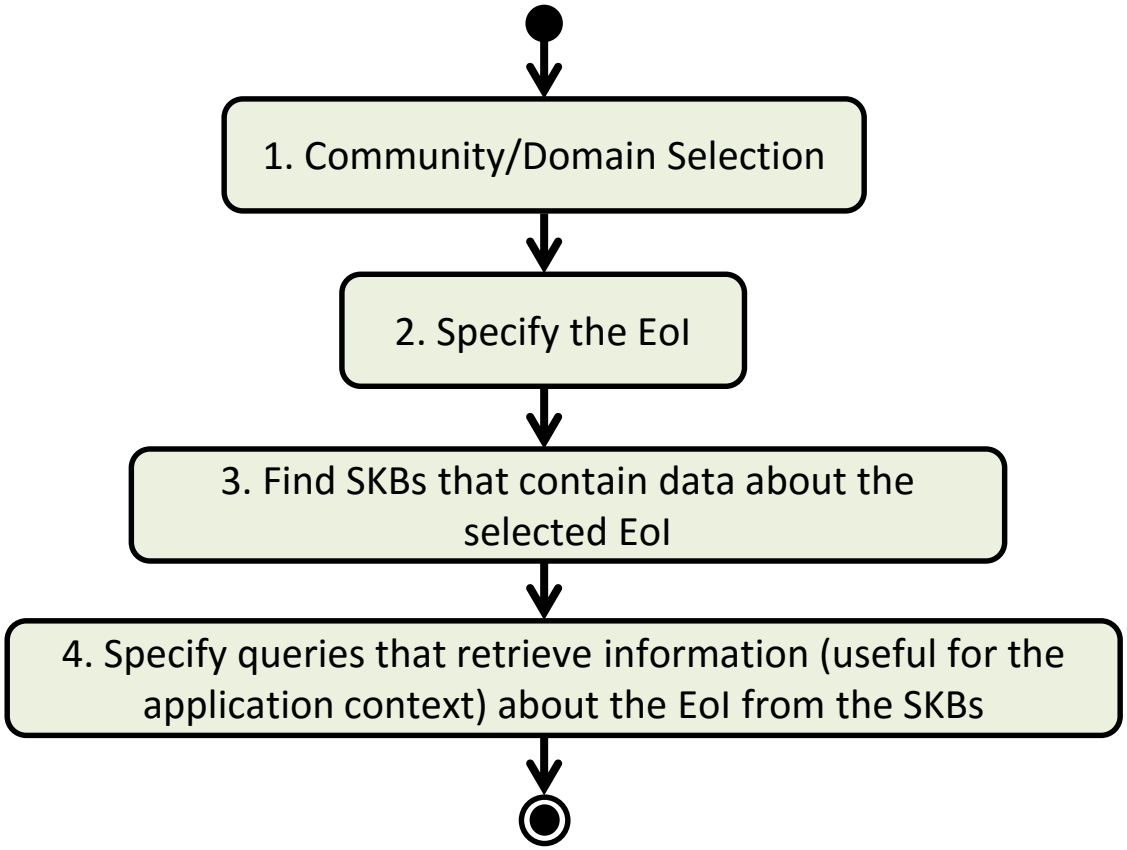
The considered Search Process

Semantic Analysis of Search Results



The considered Configuration Process

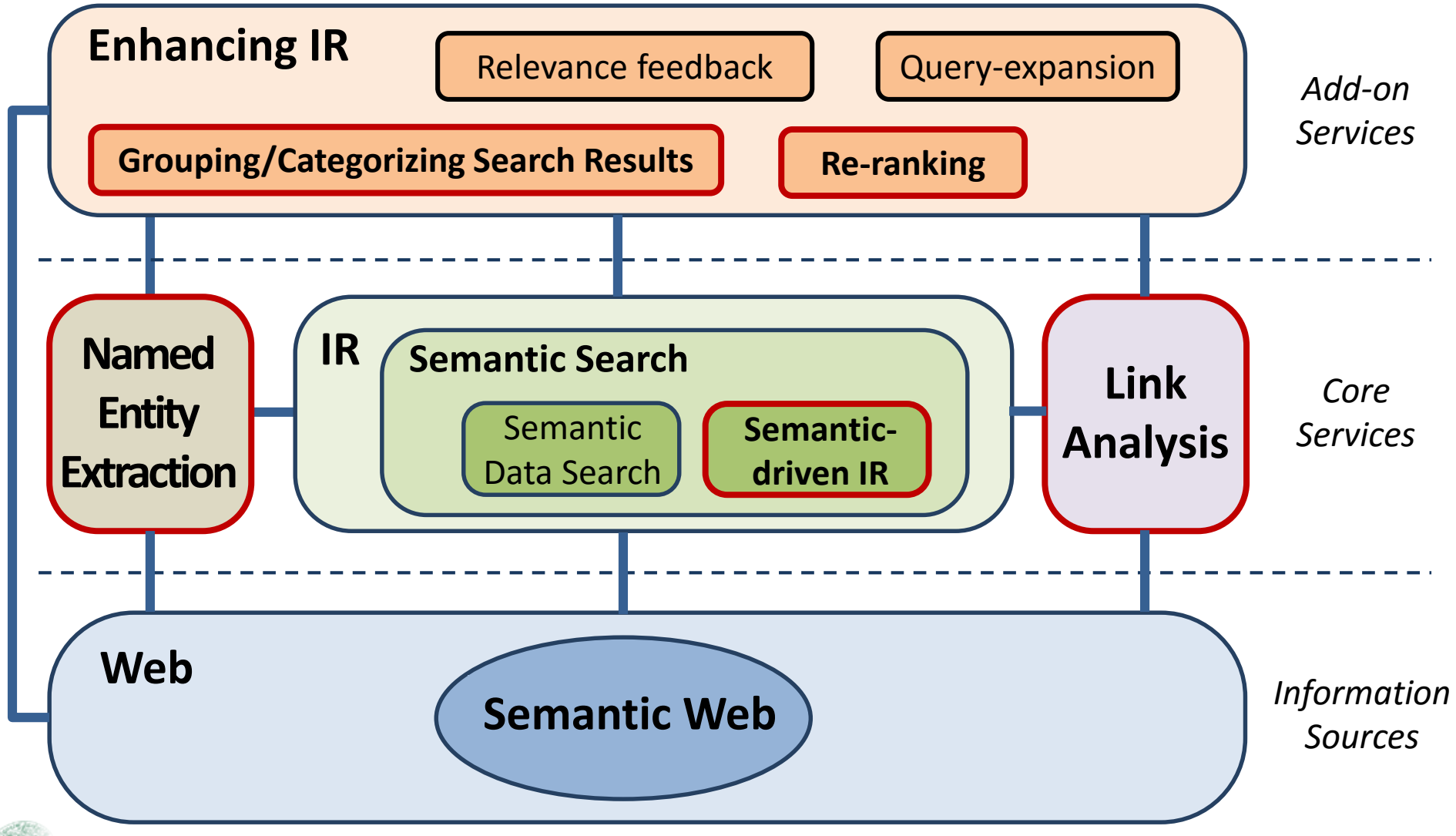




Background and Related Work

- Semantic Search
- Named Entity Extraction (NEE)
- Link Analysis

Background and Related Work



Semantic Search

- Semantic data search
  - Focus on retrieving semantic data (from a semantic corpus)
  - Swoogle, Sindice, Falcons, SWSE, Watson, ...
  - Users get as output resources, not documents
    - Quite distant from the usual way users search for information

- Semantic-driven IR
  - Understanding user's query (e.g., Google's Knowledge Graph, Yahoo's Spark)
  - Creating summaries (e.g., Microsoft's EntityCube)
  - Enriching/Constructing snippets
  - Exploiting ontologies in IR
  - **Semantic Analysis of Search Results**
    - **Entity**-based, exploiting **LOD**, at **query-time**

Named Entity Extraction (NEE)

- Identifying entities in text belonging to a set of predefined categories
  - LOD-based NEE tools: DBpedia Spotlight, AlchemyAPI, OpenCalais, AIDA, …
  - Configuring these systems is…challenging!

- Our focus: **Configurability**
  - Dynamic configuration of a NEE system (by exploiting multiple SKBs)
  - Applicable by existing NEE systems
  - Vision: Open and Exchangeable configurations (Interoperability)

Link Analysis

- Data-analysis technique used to evaluate relationships (connections) between nodes in a network or graph
- Link Analysis approaches on Semantic Data
  - Random Walk-based (PageRank-like):
    - [L. Ding et al, ISWC'05], [A. Harth et al, ISWC'09], [R. Delbru et al, ESWC'10], [A. Hogan et al, SSWS'11], [D. Mladenic et al, ESWC'12]
  - Spreading activation:
    - [C. Rocha et al, WWW'04], [M. Ciglan et al, WWW'12], [A. Agrawal et al, COMAD'12]
  - Kleinberg's HITS-like:
    - [B. Bamba and S. Mukherjea, SWDB '05], [D. Mladenic et al, ESWC'12]

- Our approach: **biased Random Walk**
  - Takes into account the **search context**
  - Exploits the **ranking** of the retrieved results for **promoting** the important (for the search context) semantic information

A. Configuring Named Entity Extraction

- A generic configuration model
- The “Open NEE Configuration Model”
- The extension of the “Open Annotation Data Model”
- The X-Link framework
- Evaluation of X-Link

Configuring Named Entity Extraction

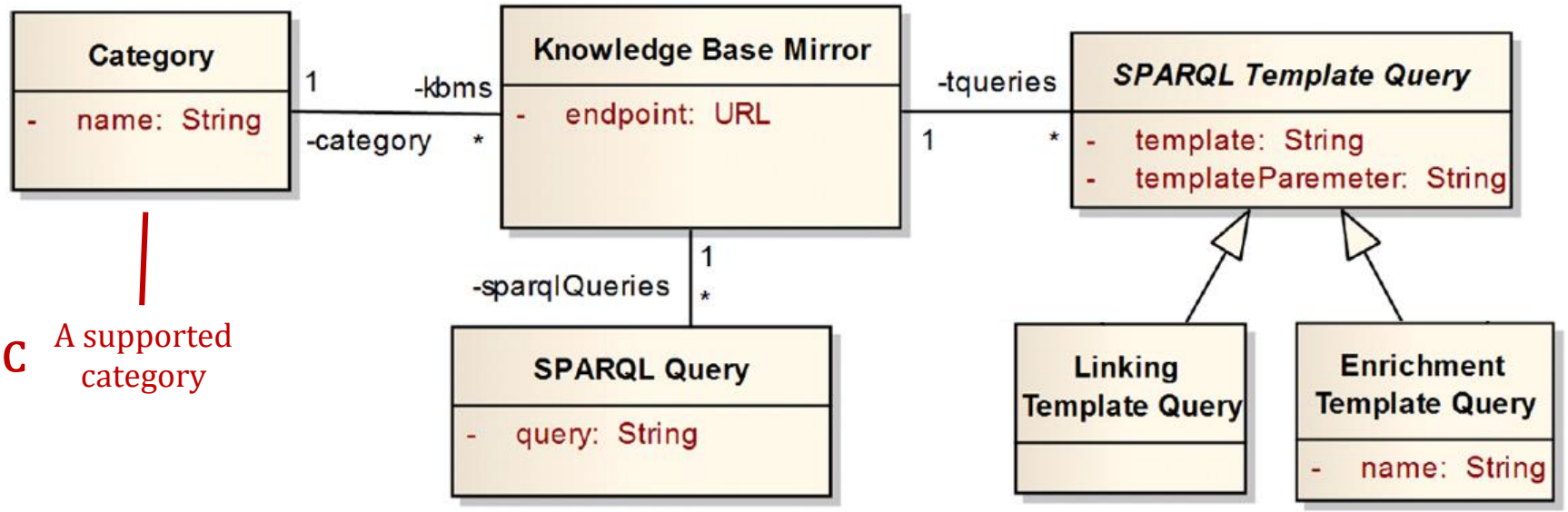
- The number of identifiable entities can be very high
- The amount of structured information that is available for these entities can be very high too

Semantic Context Specification

A

- ❑ How one can specify the "**entities of interest**" (for a given domain)?
- ❑ How to specify the **additional semantic information** about these entities that is useful to provide in a search context?
- ❑ Can we **dynamically configure** all the above and then **fetch** and **exploit** such semantic information at **query-time**?

Configuration Model



$c \in C$ A supported category

E(c) Specifying the entity names belonging to a category ***c***

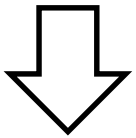
U(e) Specifying how an entity name ***e*** corresponds to entity URIs

Descr(u) Specifying what extra information (RDF statements) should be fetched for each entity URI ***u***

Specification of the Entity Names - Example

❖ Category "**Fish**" from **DBpedia** Knowledge Base

```
SELECT DISTINCT str(?label) WHERE {
   ?uri rdf:type <http://dbpedia.org/ontology/Fish> .
   ?uri rdfs:label ?label FILTER(lang(?label)='en')
}
```



http://dbpedia.org/sparql

Acanthicus
Acanthurus
Acanthurus achilles
Acanthurus albipectoralis
Acanthurus auranticavus
….

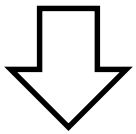
Linking Template Query – Example

❖ **Link** the detected fishes with entity URIs from **DBpedia**

```
SELECT DISTINCT ?uri WHERE {
  ?uri rdf:type <http://dbpedia.org/ontology/Fish> .
  ?uri rdfs:label ?label FILTER(regex(str(?label), '[ENTITY]', 'i'))
}
```

For the entity name "**chum salmon**":

```
SELECT DISTINCT ?uri WHERE {
  ?uri rdf:type <http://dbpedia.org/ontology/Fish> .
  ?uri rdfs:label ?label FILTER(regex(str(?label), 'chum salmon', 'i'))
}
```



http://dbpedia.org/sparql

http://dbpedia.org/resource/Chum_salmon

Enrichment Template Query – Example

❖ **Enrich** the detected fishes with properties from **DBpedia**

```
SELECT DISTINCT ?propertyName ?propertyValue
WHERE {
   <[URI]> ?propertyName ?propertyValue }
```

For the entity URI "**http://dbpedia.org/resource/Chum_salmon**":

```
SELECT DISTINCT ?propertyName ?propertyValue
WHERE {
  <http://dbpedia.org/resource/Chum_salmon>  ?propertyName ?propertyValue }
```

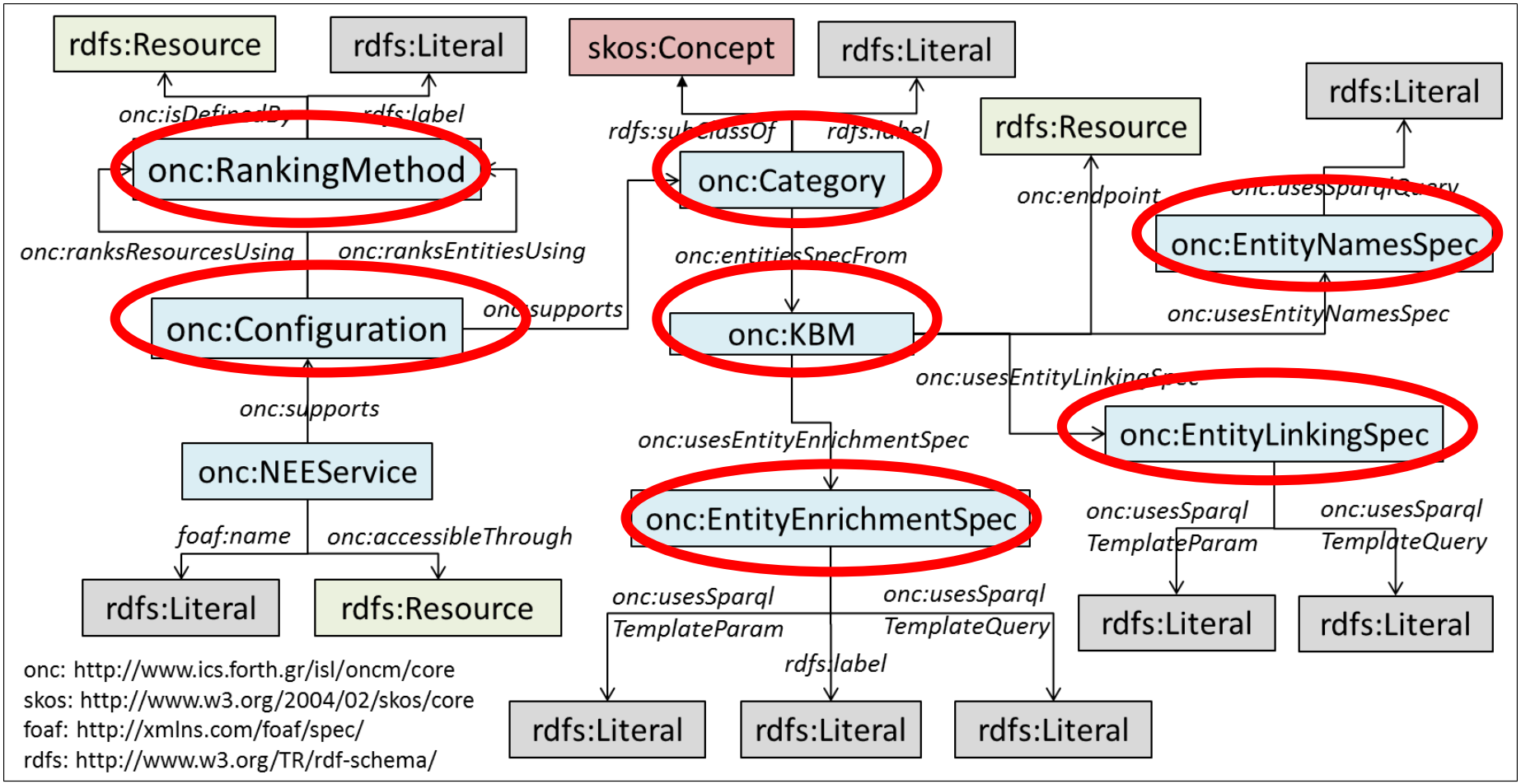
http://dbpedia.org/sparql

http://dbpedia.org/ontology/**binomialAuthority**
http://dbpedia.org/ontology/**class**
http://dbpedia.org/ontology/**family**
http://dbpedia.org/ontology/**genus**
… … … …

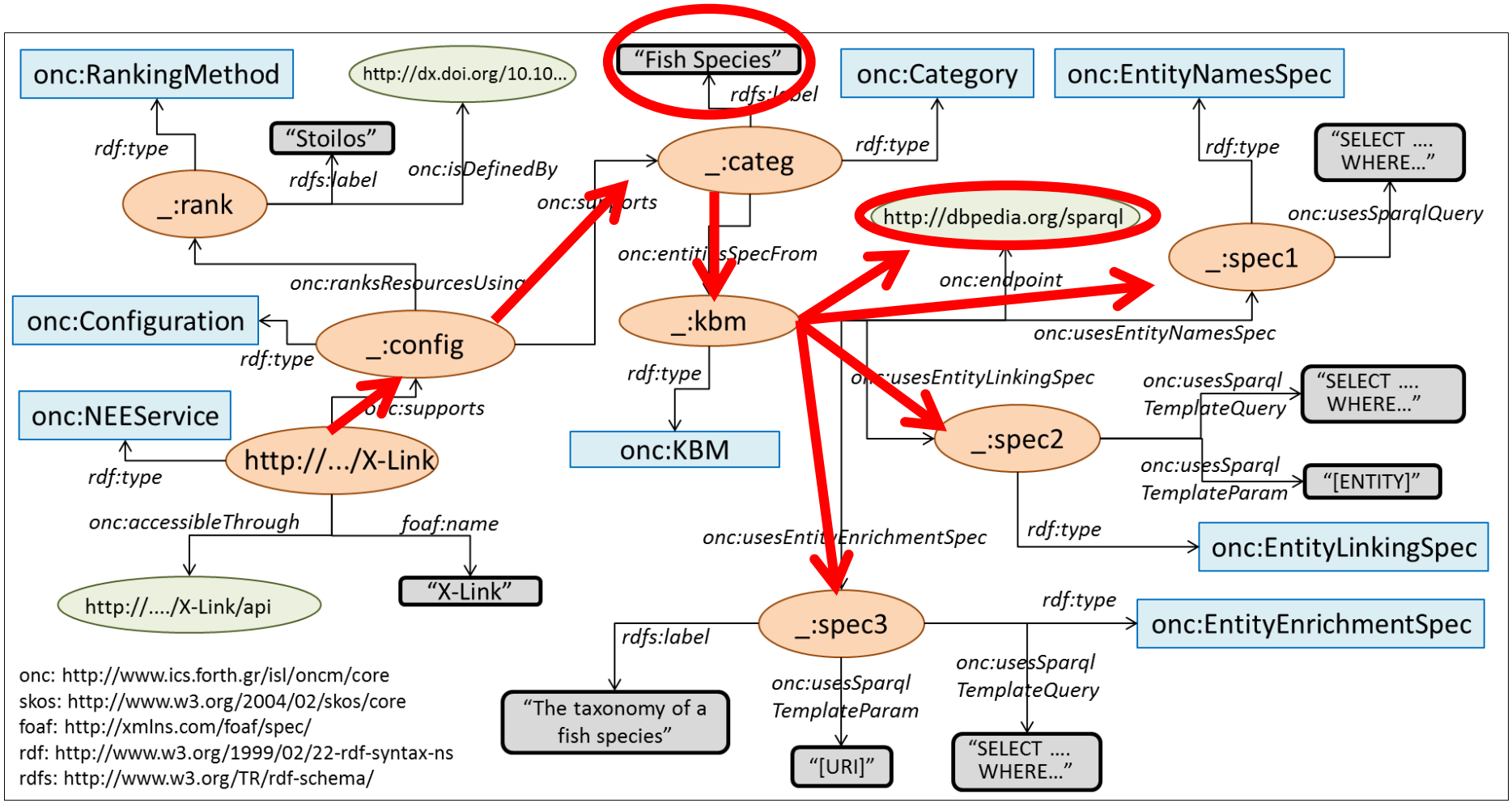
http://dbpedia.org/resource/**Johann_Julius_Walbaum**
http://dbpedia.org/resource/**Actinopterygii**
http://dbpedia.org/resource/**Salmonidae**
http://dbpedia.org/resource/**Oncorhynchus**
… … … …

The “Open NEE Configuration Model”

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

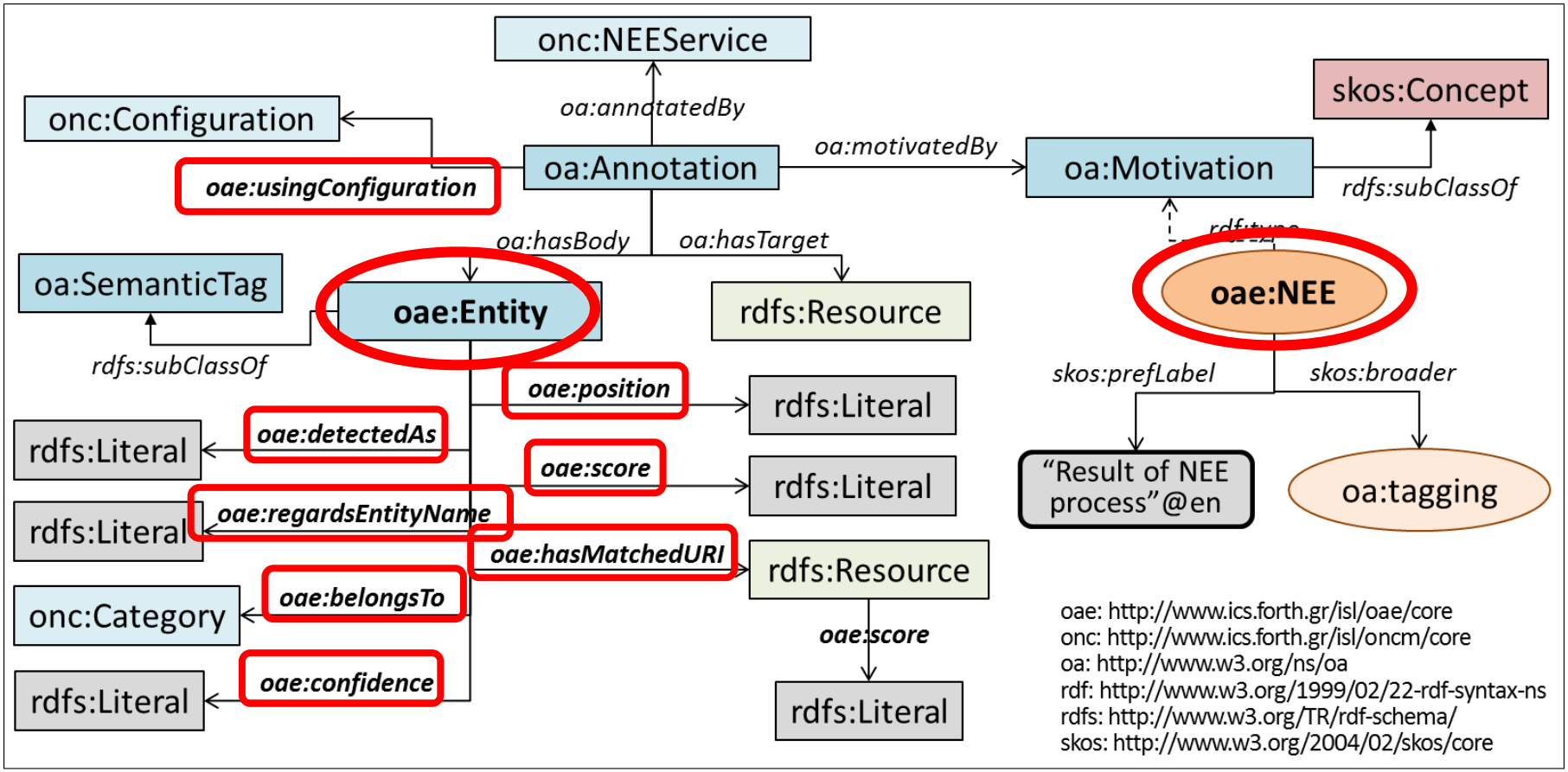


Instantiation Example

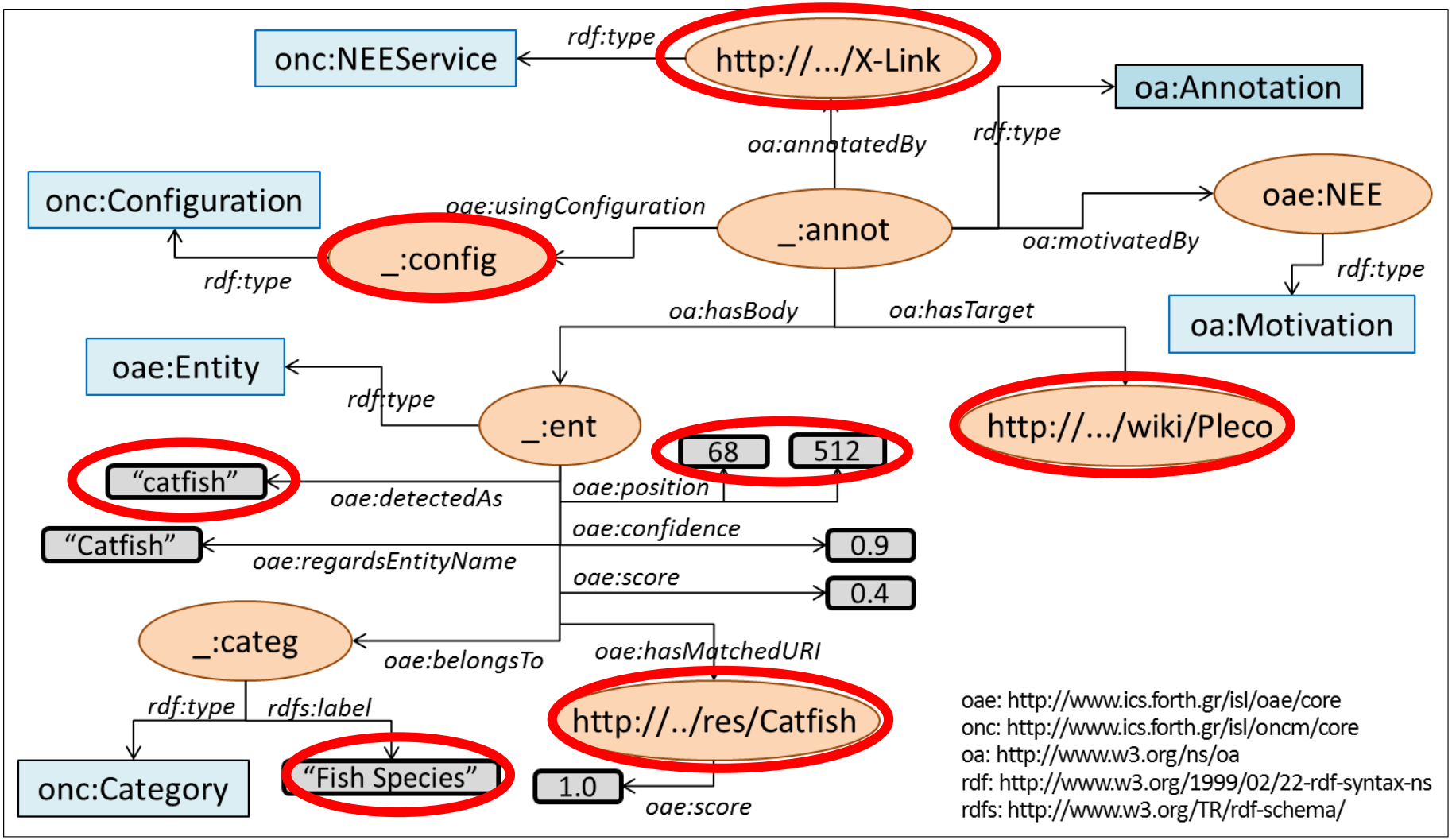


Open Annotation Extension

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



Instantiation Example



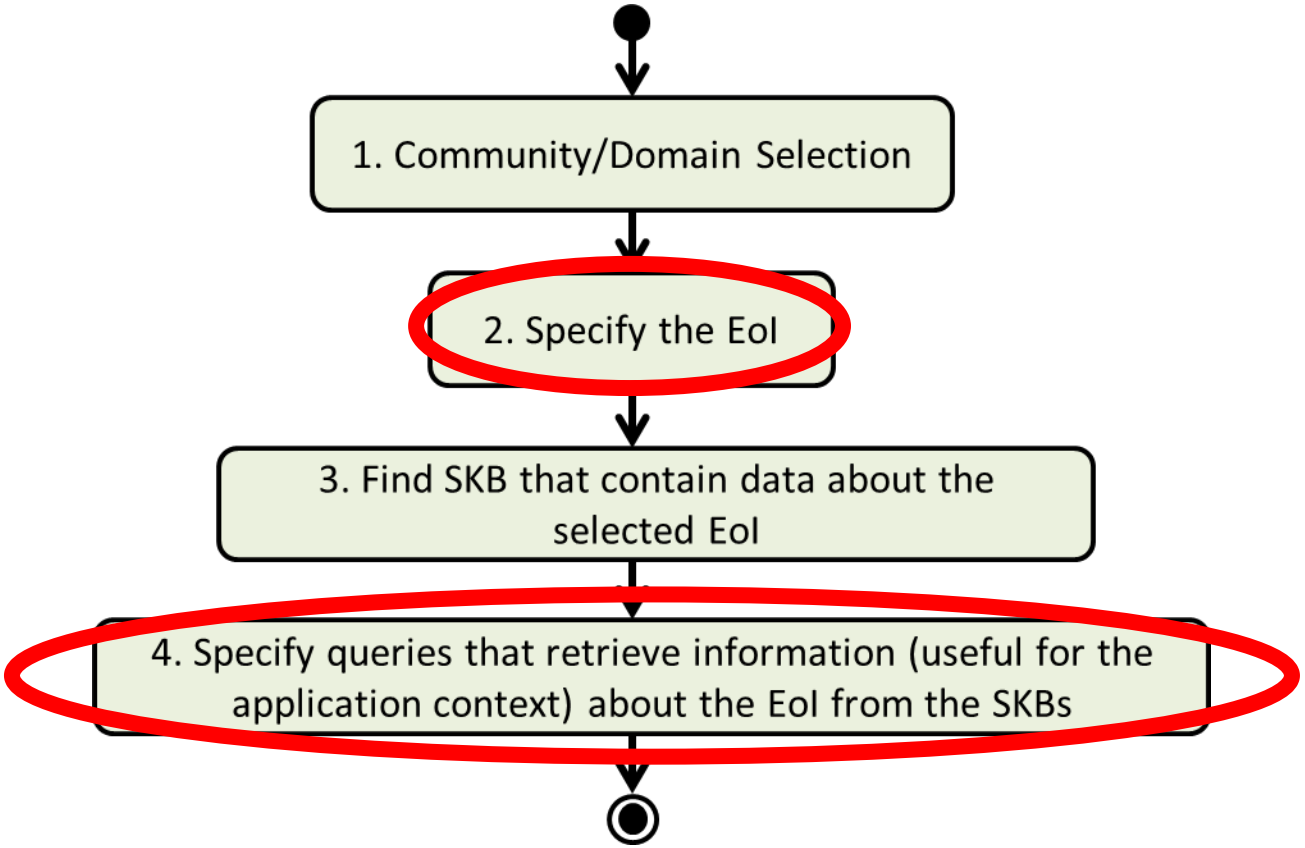
The **X-Link** framework

http://www.ics.forth.gr/X-Link

- NEE **framework** that adopts the proposed configuration model
- Entity Mining using Gate ANNIE
  - Supported file types: Plain text, HTML, PDF, .doc, .docx, .ppt, .pptx, XML-based
  - Output: XML, CSV, RDF (exploiting the **Open Annotation Extension**)
- Entity **Linking** and Entity **Enrichment**
  - according to the specified **SPARQL template queries**
- Infer the **Connectivity** of the identified entities
  - Return the RDF statements that connect one or more of the identified entities
- Entity **Disambiguation**
  - Master's Thesis of Manolis Baritakis (2016)
- **Configurability**
  - Can read a configuration expressed in RDF using the **Open NEE Configuration Model**
  - Java API

X-Link Evaluation

- **Task-based User Study**: Ease of configuration
- **Case Study**: Efficiency



Task-based User Study

- Objective
  - Test the **ease of configuring** X-Link
  - Identify usability problems
- Target User
  - an **admin** or **developer** who wants to use X-Link for building and dynamically configuring an application

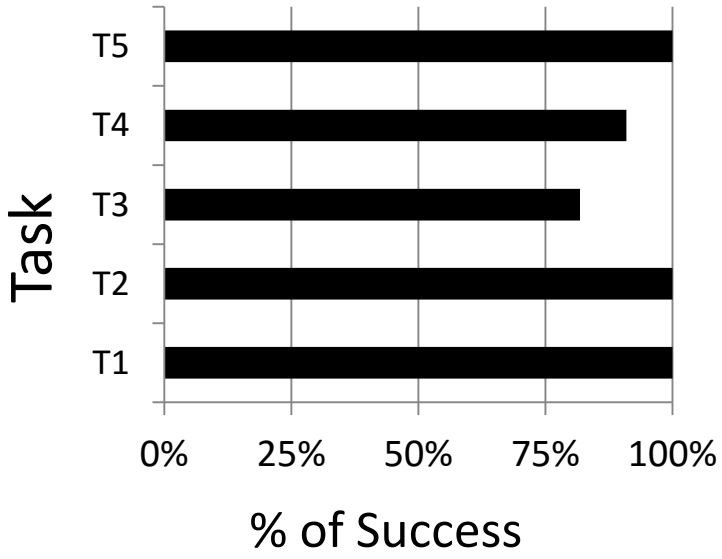
- X-Link was deployed as a Web Application configured for the **marine domain**
  - Identification of **Fish Species** in a text or Web page
- The admin can change the configuration through an admin page
  - **Add** and **update** categories, specify how to **link** and **enrich** the entities

Task-based User Study

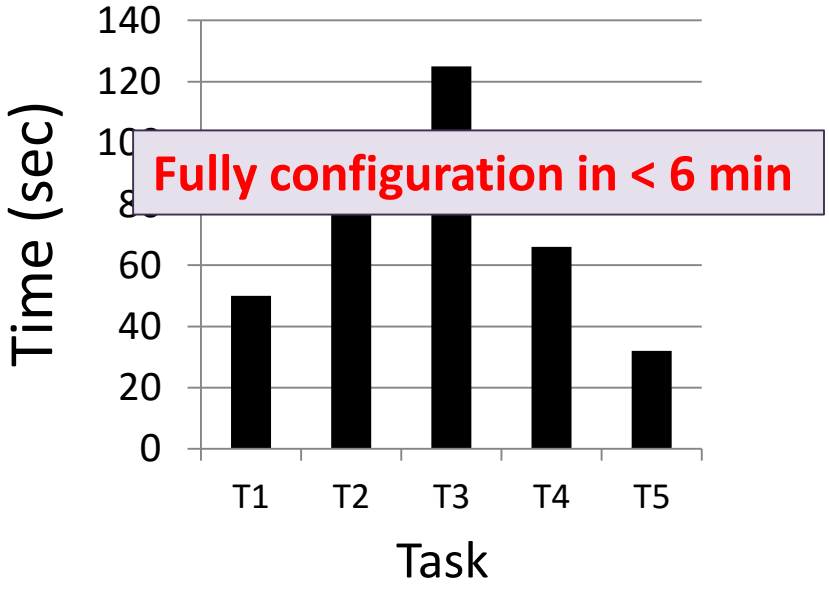
T1	Add the category “European Countries”
T2	Update the category Fish with fish names in Spanish
T3	Link the identified Fishes with resources from DBpedia
T4	Enrich the identified Fishes with their outgoing properties
T5	Apply entity mining and inspect connectivity

Results from 11 users

Success rate for each task



Average time for completing successfully each task



Case Study: Querying online DBpedia

- Objective
  - Examine feasibility and efficiency
- We measured the time (including network delay time) for:
  - **creating** a new category
    - Using 35 resource classes (as categories of entities)
  - **linking** an identified entity with semantic resources
    - Using 40 resource classes (as categories of entities) and 10 randomly selected labels for each class
  - **enriching** an entity URI
    - Using 160 randomly selected entity URIs and several types of retrieved properties
- The results highly depend on the efficiency of the queried endpoint at the time we run the experiments!
  - The results may vary if we query another endpoint, or at another moment

Case Study: Querying online DBpedia

- **Creating** a new category
  - Time depends on the number of entities returned
  - 10,000 entities → 15 sec | 100,000 entities → about 5 min
  - Limitations by the endpoint on the number of returned results per request!
- **Linking** an entity
  - Time depends on the number of entities in the category
  - 100,000 entities → about 1 sec | 1m entities → about 5 secs
- **Enriching** an entity
  - Time depends on the type of properties
  - Very low time, less than 0.5 sec

Lessons Learned – **Reliability** and **Scalability**

- Existing publicly available knowledge bases (like DBpedia) are not reliable
  - They mainly serve demonstration purposes
  - Their efficiency and availability change over time
- For being reliable, a dedicated Warehouse can be constructed that will only serve our application
  - Distributed infrastructure

B. Stochastic Ranking of Entities and Properties

- Semantically-Enriched Graph of Identified Entities (SEGIE)
- State Transition Graph (STG)
- Analyzing the STG
- Promoting the Important Entities
- Top-K Semantic Graphs
- Evaluation

Stochastic Ranking of Entities and Properties

- The number of entities identified in the search results can be high
- The amount of semantic information that has been retrieved for these entities can be very high too (their associations and properties)

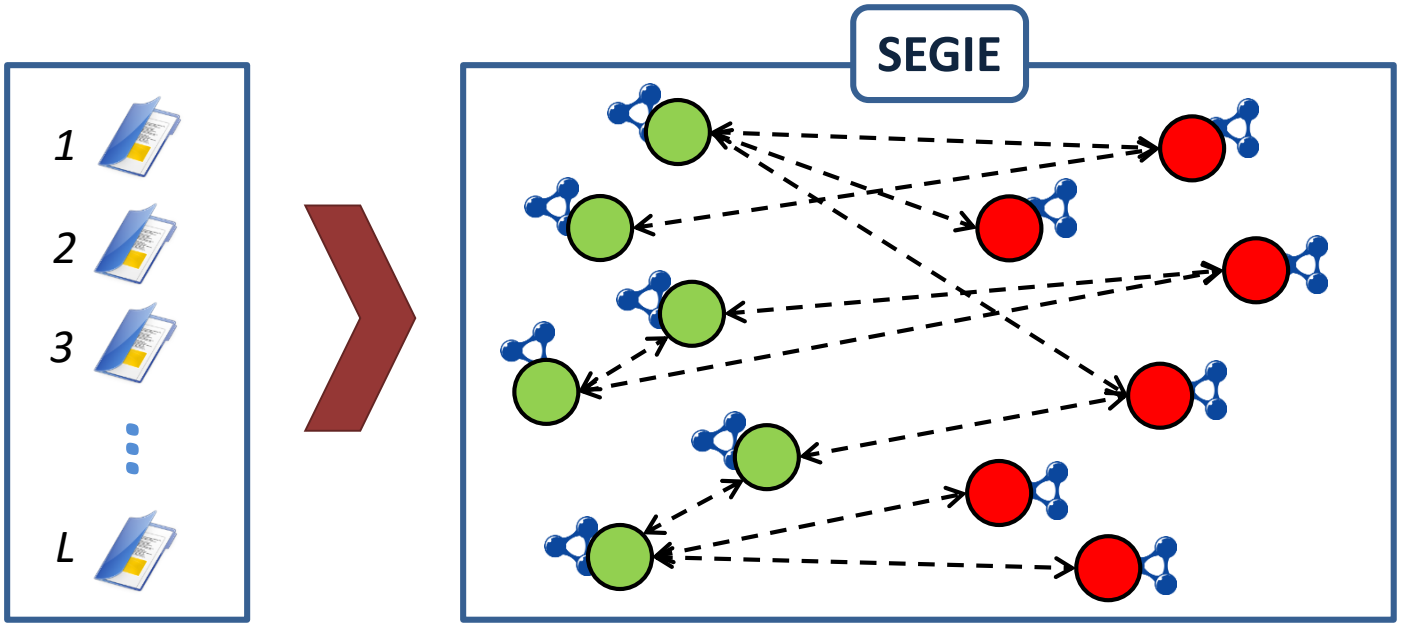
Semantic Context Selection

B

- ❑ How to **select** the entities (as well as entity associations and properties) that better characterize the search results?
- ❑ Can this analysis and selection be done **at query-time**?

Stochastic Ranking of Entities and Properties

- Semantically-Enriched Graph of Identified Entities (SEGIE)



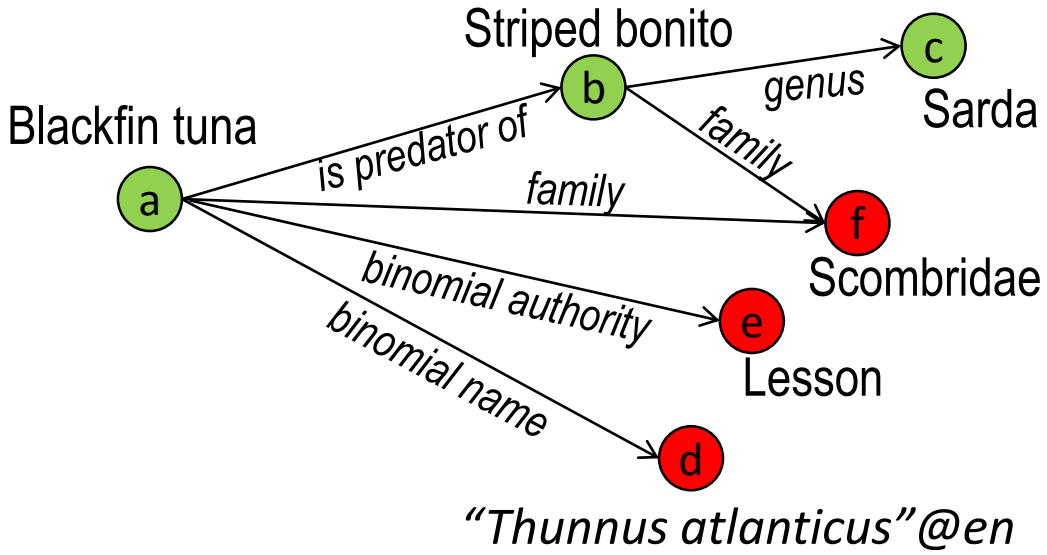
Top-L Search Results

LOD Entities identified in the retrieved results (result of NEE)

Properties, related entities and associations (result of Entity Enrichment)

Stochastic Ranking of Entities and Properties

- Semantically-Enriched Graph of Identified Entities (SEGIE)

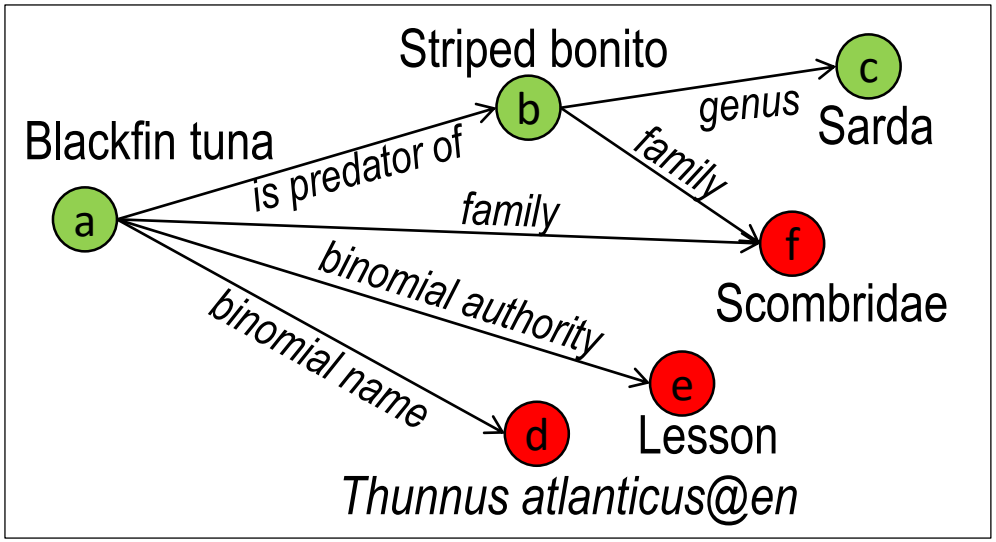


Identified in the search results (set E) – Result of entity mining

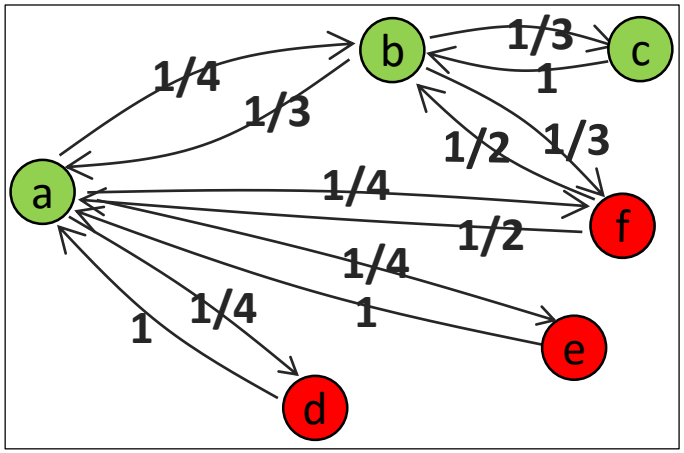
Retrieved from LOD (set R) – Result of entity enrichment

The State Transition Graph

SEGIE

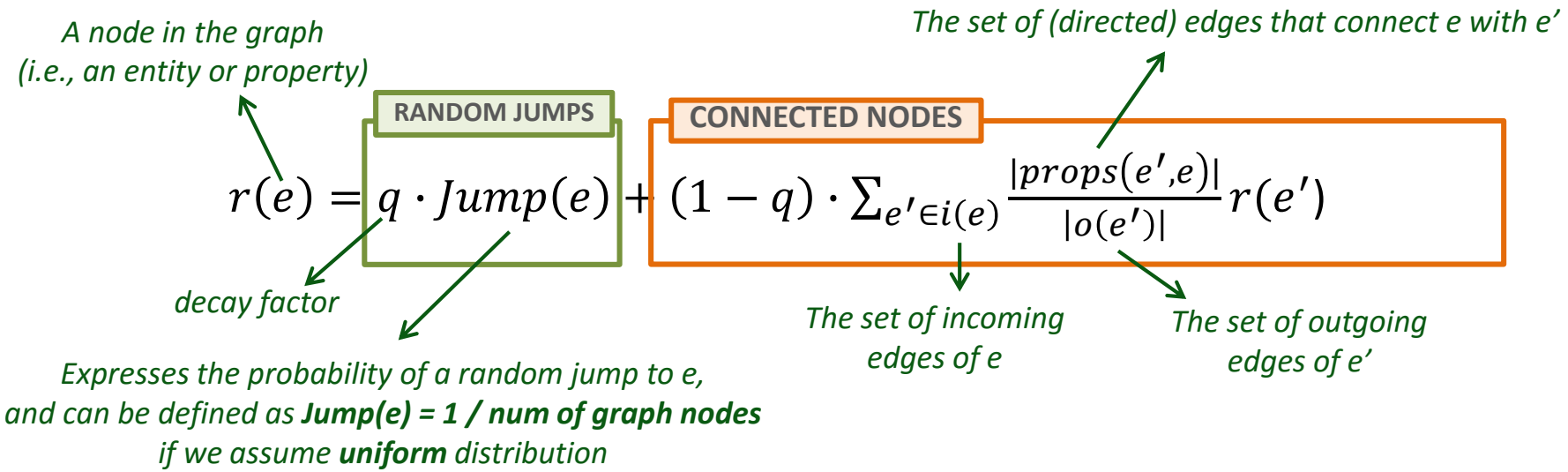


State Transition Graph (STG)



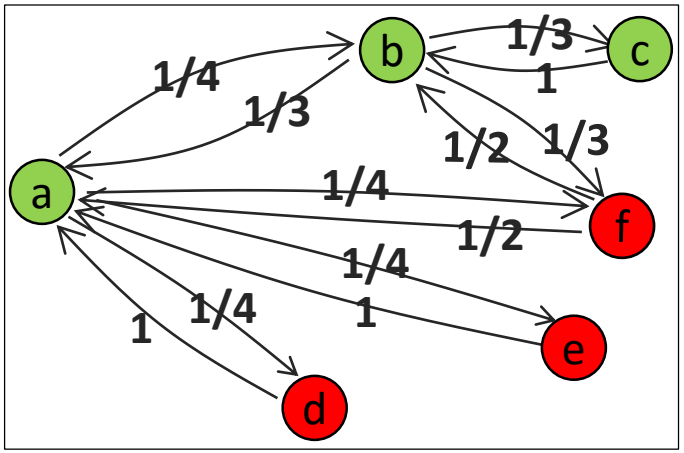
Analyzing the State Transition Graph

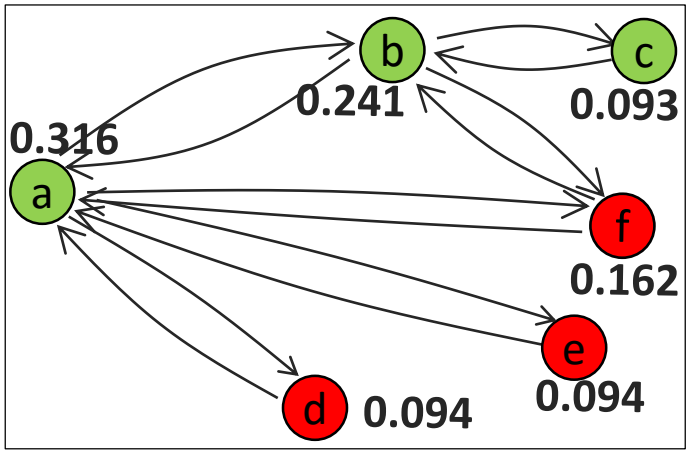
- Apply a probabilistic (PageRank-like) algorithm for identifying the important entities and properties



- The scores are computed iteratively and iterations should be run to convergence

Analyzing the State Transition Graph





- a Blackfin tuna ≈ 0.316
- b Striped Bonito ≈ 0.241
- f Scombridae ≈ 0.162
- d e Lesson, Thunnus atlanticus@en ≈ 0.094
- c Sarda ≈ 0.093

- *Uniform distribution for the random jumps*
- *Decay factor = 0.15, Performing 10 iterations*

Promoting the Important Entities

- The "importance" of the entity "e" in document "a"

$$imp(e,a) = \sum_{p \in P} \left(\frac{ef(e, a_p)}{max_{e' \in ent(a_p)} ef(e', a_p)} w(p) \right)$$

Frequency (num of occurrences) of the entity "e" in the part "p" of doc "a"

normalization

Document parts (e.g. title, abstract, etc.)

The weight of a document part p∈P (expressing its importance)

- The importance of the entity "e" in the set of top-L documents (results)

$$HitScore(e) = \sum_{a \in docs(e)} (imp(e,a)\ score(a))$$

Docs containing the entity "e"

the score of doc "a" as returned by the search system

The score is higher if the entity has been identified in the top-scored documents

Promoting the Important Entities

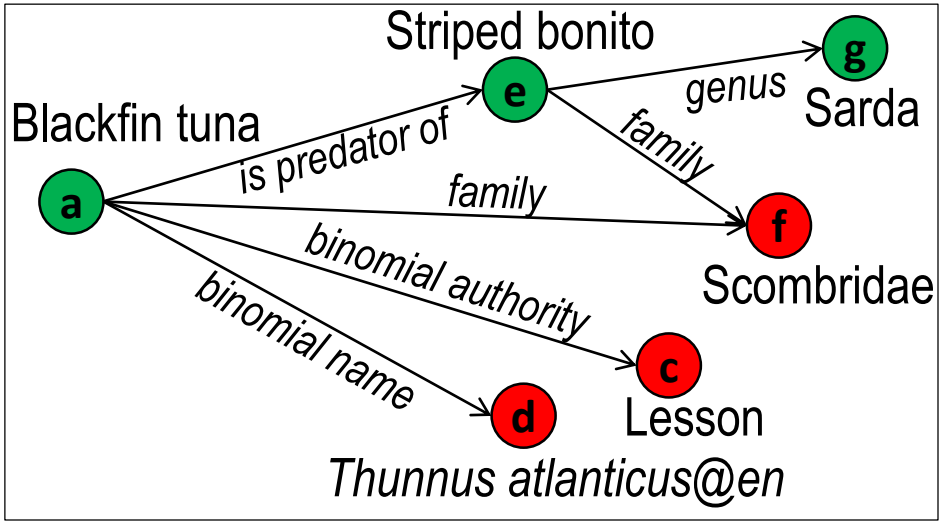
- We **bias** the random jumps

$$r(e) = q \cdot \boldsymbol{Jump(e)} + (1-q) \cdot \sum_{e' \in i(e)} \frac{props(e',e)}{|o(e')|} r(e')$$

$$Jump(e) = \frac{HitScore(e)}{\sum_{e' \in E} HitScore(e')}$$

- The probability of a random jump to an entity is higher if the entity has been identified in the top-scored documents
- The probability of a random jump to entities/properties **not** identified in the retrieved documents is zero

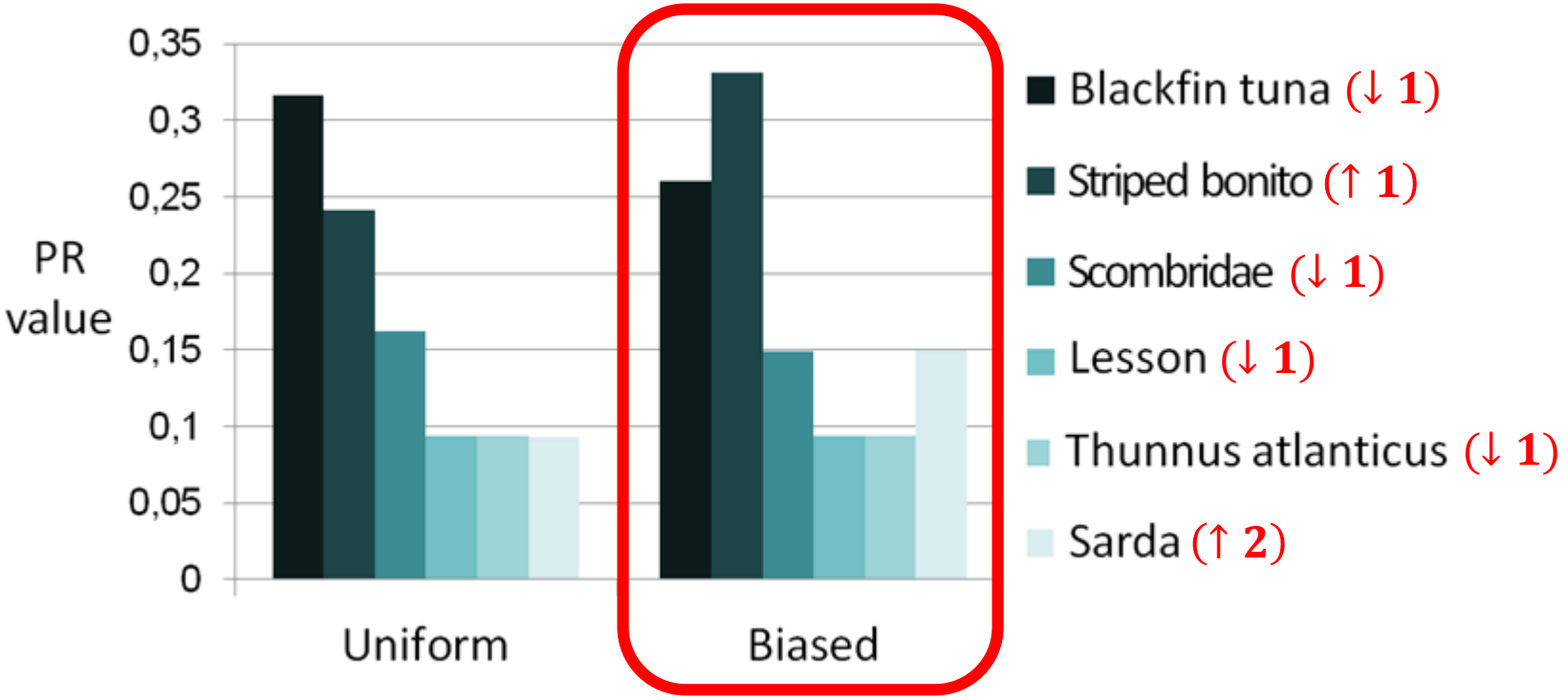
Promoting the Important Entities



We perform entity mining in the top-10 results and get the following results:

- **Striped bonito** was detected in the **1st**, **2nd** and **3rd** result
- **Sarda** was detected in the **1st** and **3rd** result
- **Blackfin tuna** was detected in the **8th** result only

Promoting the Important Entities

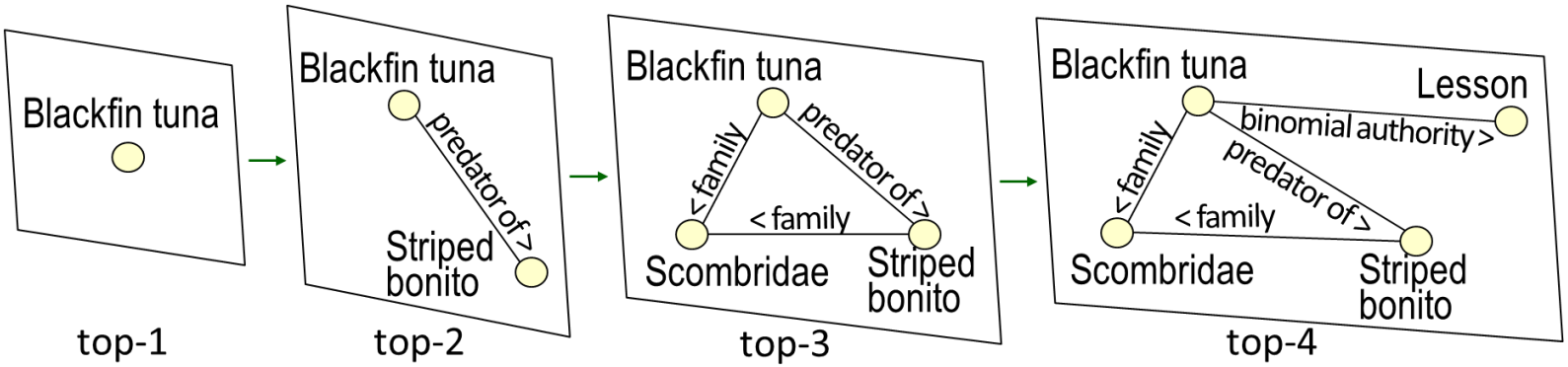


1. **Blackfin tuna**
2. **Striped Bonito**
3. Scombridae
4. Lesson | Thunnus atlanticus@en
5. **Sarda**

1. **Striped Bonito**
2. **Blackfin tuna**
3. **Sarda**
4. Scombridae
5. Lesson | Thunnus atlanticus@en

Top-K Semantic Graphs

- The system can return and show to the user:
  - A top-K list related to the search results (containing the top-K entities and properties)
  - A top-K graph related to the search results
    - Vertices: the K most highly-score nodes
    - Edges: the edges that connect these top-K nodes



- A survey in **Marine Search** has shown that the majority of users would prefer to see a graph representation of the identified entities [Fafalios et al., 2014]
  - However, visualization is challenging (especially for big K)

Effectiveness

- Objective
  - Evaluate if the proposed ranking scheme produces more preferred rankings compared to other approaches
- User study in the **marine** domain
  - No standard collection or evaluation procedure!

- Comparative evaluation of:
  - Proposed (biased) PageRank-like algorithm (**BiPR**)
  - Plain (no-biased) PageRank algorithm (**PR**)
  - Spreading Activation (**SA**)
    - Initial activation = normalized HitScore
    - α = 0.85
    - firing threshold = very small real number

- decay factor = 0.15
- 50 iterations

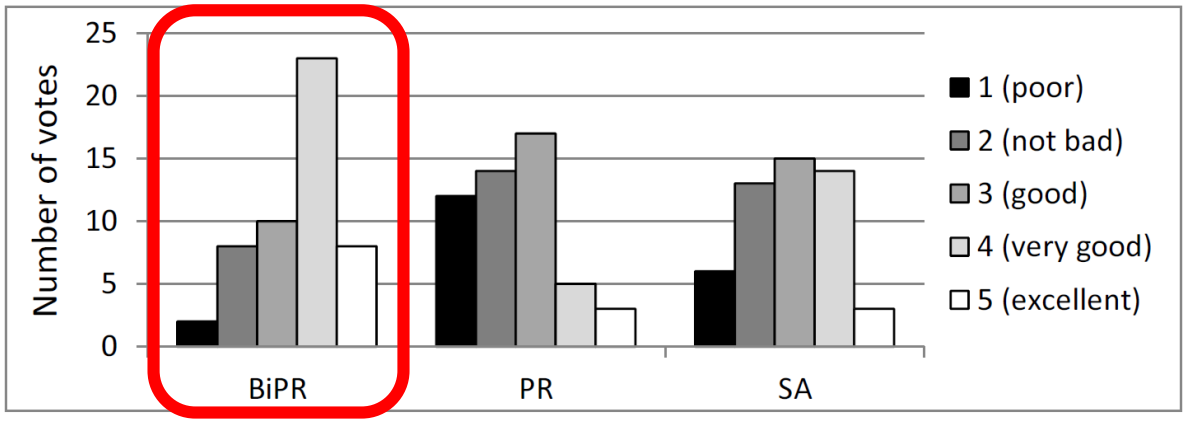
Effectiveness

- We deployed a **Web Search** application which implements the proposed functionality
  - Keyword-based queries
  - Entity mining in the **top-100 snippets** returned by **Bing**
  - **Fish Species** (from DBpedia) as the entities of interest
  - **DBpedia** for retrieving the incoming and outgoing properties of the identified entities
- The users could submit their own queries

- For each submitted query, three **top-10 lists** of (ranked) entities and properties are displayed
  - the one next to the other with random display order
  - each one is produced by one of the three ranking schemes (BiPR, PR and SA)
- The user can evaluate each ranking by selecting one of the following options:
  - **1** (poor), **2** (not bad), **3** (good), **4** (very good), **5** (excellent)

Effectiveness

- 51 queries were submitted, from 17 participants (mainly marine biologists)
- For each query in average:
  - **12** entities detected in the search results
  - **4,685** triples were derived from DBpedia
  - **2,031 entities and properties had to be ranked**



Average scores:

- Biased PageRank (BiPR): average: **3.53/5** (good to very good), median: **4** (very good)
- Spreading Activation (SA): **2.9/5** (almost good), median: **3** (good)
- PageRank (PR): **2.47/5** (not bad to good), median **2** (not bad)

Efficiency of Semantic Analysis

- Time for NEE depends on many parameters
  - *Number of top results we want to analyze, Text size of each result/document, Efficiency of the underlying NEE algorithm, …*
- Real-time NEE in the top-100 results returned by a Patent Search System *(using X-Link and supporting 4 categories of entities):*
  - ≈1 sec [Fafalios et al. 2013]

- We measure the time for:
  1. Creating the SEGIE (accessing DBpedia's online Turtle files at real-time)
  2. Creating the STG
  3. Running PageRank
  4. Creating a top-500 graph
- Experiments for several numbers of randomly selected entities belonging to 10 randomly selected RDF classes

- The experiments were carried out using an ordinary laptop with processor Intel Core i5 @ 2.4Ghz CPU, 4GB RAM and running Windows 7 (64 bit). The implementation is in Java 1.7 and for the creation and the management of the graphs we use the Java Universal Network/Graph Framework (JUNG). For achieving accuracy we repeated the experiments 20 times.

Efficiency

#entities	SEGIE #vertices	SEGIE #edges	STG #edges	Top-500 Graph #edges
50	2,573	3,790	7,580	889
100	4,133	6,193	12,386	1,493
500	20,743	34,816	69,632	3,471
1,000	49,954	84,893	169,786	3,411
10,000	528,815	995,981	1,991,962	3,421

#entities	SEGIE creation time	STG creation time	Time for Running PageRank	Top-500 creation time
50	1.4 sec	28 ms	194 ms	42 ms
100	2.9 sec	95 ms	329 ms	68 ms
500	13 sec	298 ms	1.7 sec	343 ms
1,000	27 sec	480 ms	3.9 sec	552 ms
10,000	258 sec	8 sec	58 sec	22 sec

- Bounding the maximum response time
  - Retrieve LOD only for the top-m (e.g., m = 100) entities (according to **HitScore**)

C. Stochastic Re-Ranking of Search Results

- Semantic Graph of Documents and Entities
- Modeling a Random Walker
- Analyzing the graph
- Evaluation

Stochastic Re-Ranking of Search Results

- Ranking is always important
  - Even when the user restricts the search space to a few results by selecting an identified entity, we must decide how to rank these results
- Classic IR techniques often do not work well in more complex information needs

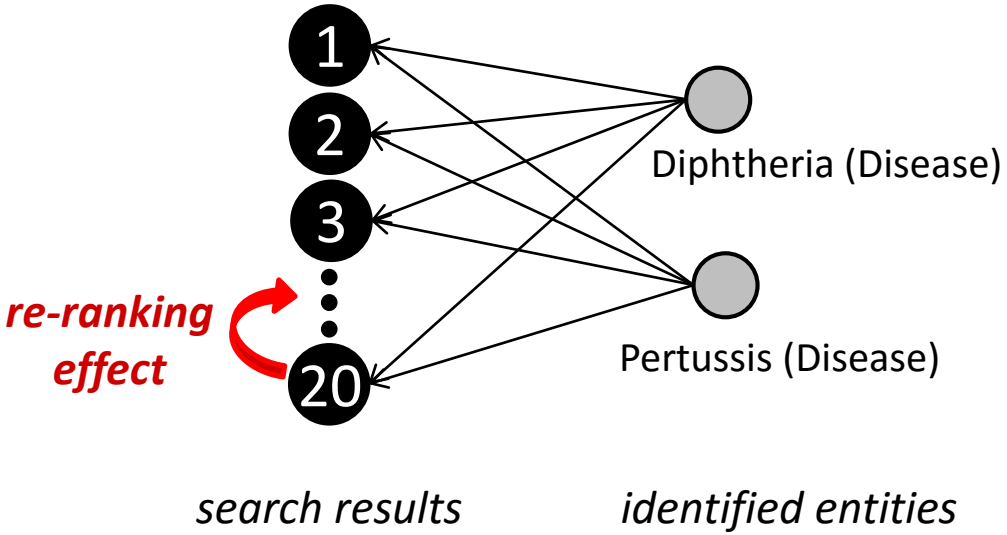
Semantic Context Exploitation

C

❑ Can we exploit the extracted entities (and the related semantic information) also in pure IR (e.g., for producing a better ranking for the search results)?

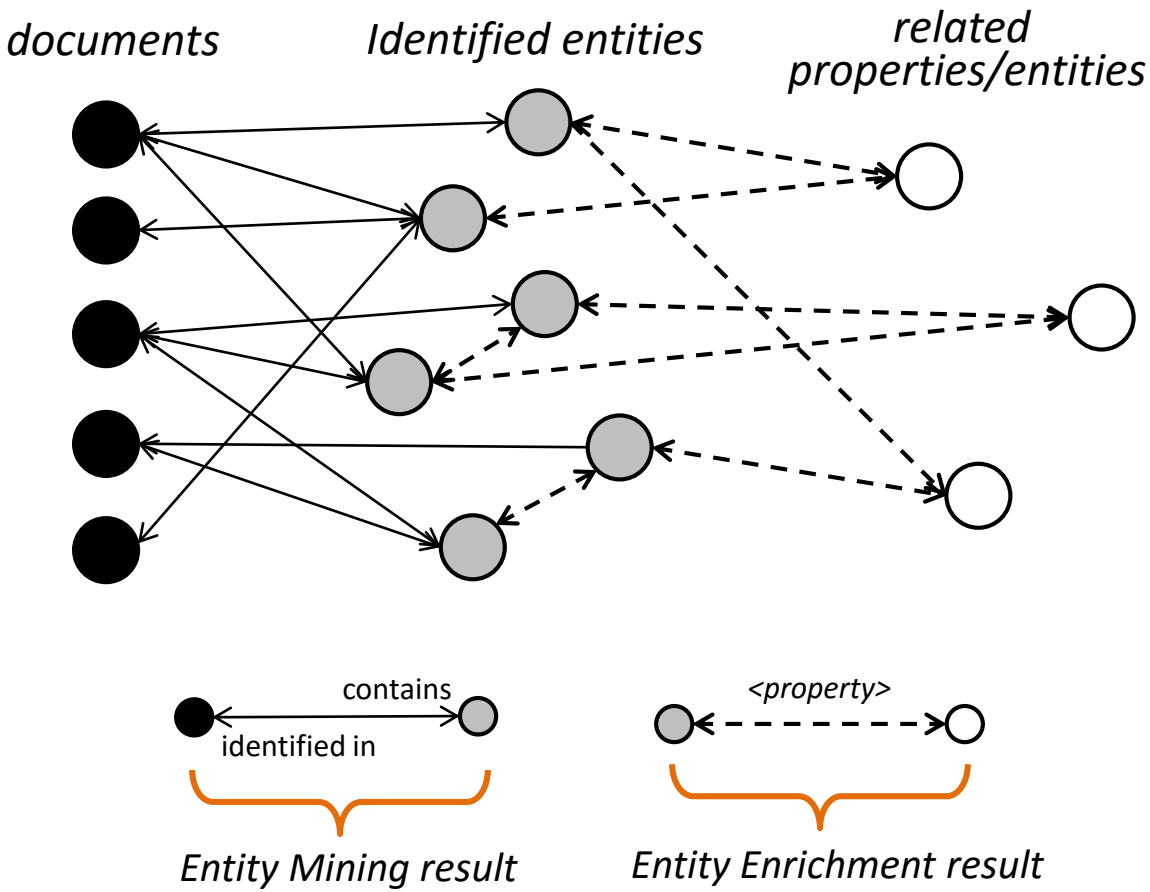
Stochastic Re-Ranking of Search Results

- Intuition:



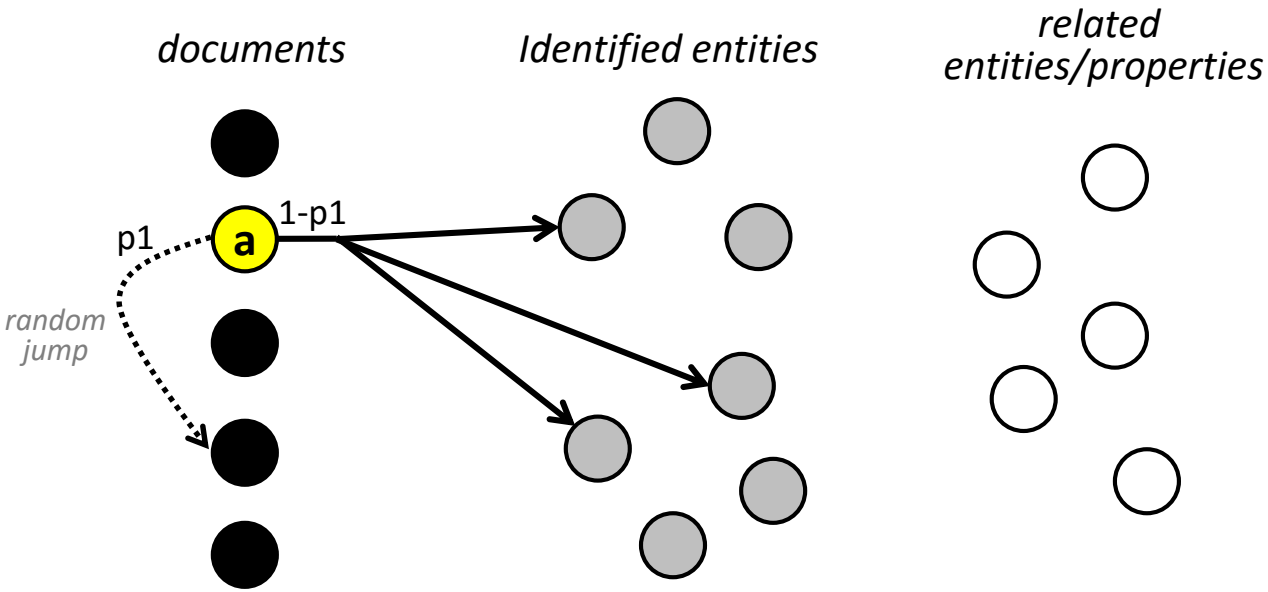
Semantic Graph of Docs and Entities

(transformed to a **State Transition Graph**)



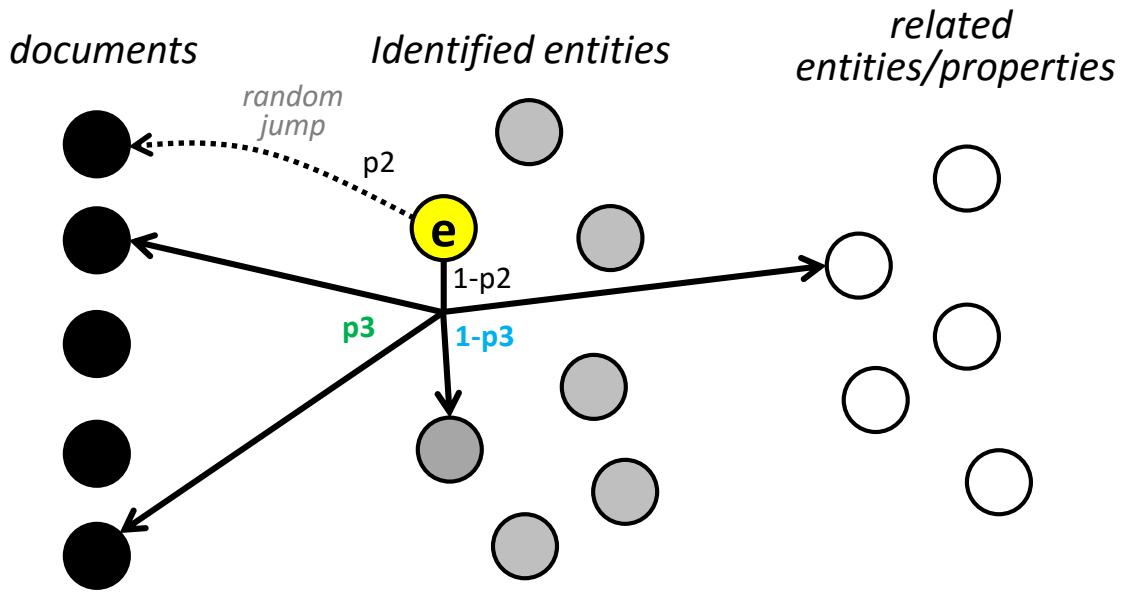
Modeling a **Random Walker**

- Whenever the walker is at a document-node "a":
  a) With probability "p1" he jumps to another hit (based on **doc scores/ranks**: the higher the score/rank of a hit is, the higher is the probability to jump to that hit)
  b) With probability "1-p1" he moves to a connected entity-node "e" corresponding to an entity mined from "a" (based on "**entity importance**" scores, i.e. **HitScore**: the higher the score of an entity is, the higher is the probability to move to that entity)



Modeling a **Random Walker**

- Whenever the walker is at an entity-node "e":
  - c) With probability "p2" he jumps to a hit (based on **doc scores/ranks**)
  - d) With probability "1-p2" he follows an edge from "e", specifically:
    - e) With probability "p3" he moves to a doc that contains "e" (based on **doc scores/ranks**)
    - f) With probability "1-p3" he moves to a connected entity/property (**equiprobably**)



Analyzing the Graph

- Apply a PageRank-like algorithm for scoring the graph nodes

$$r(n) = d \cdot Jump(n) + (1-d) \cdot \sum_{n' \in in(n)} (weight(n' \rightarrow n) \cdot r(n'))$$

Random jump probability
(p1 = p2 = d)

$$Jump(n) = \begin{cases} 0 & n \notin A \\ \frac{score(n)}{\sum_{a' \in A} score(a')} & n \in A \end{cases}$$

$$weight(n' \rightarrow n) = \begin{cases} \frac{HitScore(n)}{\sum_{e' \in ent(n')} HitScore(e')} & n' \in A,\ n \in E \\ p_3 \cdot \frac{score(n)}{\sum_{a' \in A} score(a')} & n' \in E,\ n \in A \\ (1-p_3) \cdot \frac{1}{|out(n')|} & n' \in E,\ n \notin A \\ \frac{1}{|out(n')|} & n' \in R,\ n \in E \end{cases}$$

Documents containing highly-scored entities will receive a high PageRank score

Effectiveness of Results Re-ranking

- 2014 and 2015 datasets of **TREC Clinical Decision Support Track**
- Indexing and Searching using **Apache Lucene** (baseline)
  - Using its **Standard Analyzer** (stemming, downcasing, filtering out stopwords) and the **default Query Evaluator** (Boolean + VSM)
- 60 topics (medical case narratives)
- Query = **topic description**
- Entity Mining using **X-Link**
  - Entities of Interest: **diseases**, **drugs**, **proteins**, **chemical substances**
  - Knowledge Base: **DBpedia**

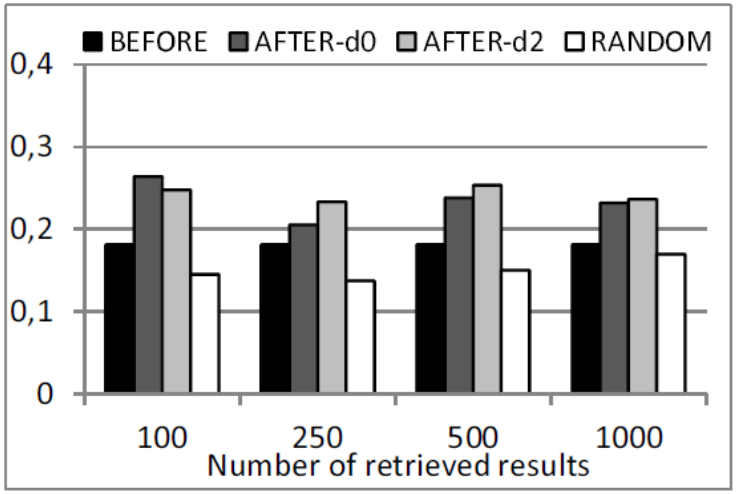
Effectiveness of Results Re-ranking

- Experiments
  - For different number of retrieved results (**L**): **100, 250, 500, 1000**
  - For different random jump (**d**) probabilities: **0.0, 0.2**
  - With/without entity enrichment
- Comparison:
  - Initial top-100 list **vs.** top-100 list after re-ranking **vs.** random top-100 list
  - Evaluation using 6 evaluation metrics
    - Bpref, AveP', nDCG', Q-Measure, rpref_relative_2, P10'
    - They are based on condensed lists (removing unjudged documents)

- Lucene did not perform well
  - ~11 relevant hits in average (in the top-100 returned list)
  - ~36 non-relevant hits in average (in the top-100 returned list)
  - ~53 unjudged hits in average (in the top-100 returned list)

Effectiveness of Results Re-ranking

- Re-ranking effect (no entity-enrichment)
  - Improved top-100 lists for the majority of topics (for all metrics and all number of retrieved hits)
- For example, for **d = 0.2** and **L = 500**:
  - **40%** better Bpref, **21%** better AveP', **24%** better nDCG'
  - Increased relevant-for-sure hits for **43/60** topics (+1 for 13 topics, +2 for 6 topics, + >2 for **24** topics)
  - Decreased relevant-for-sure hits for **7/60** topics (-1 for **5** topics, -2 for 1 topic, -3 for 1 topic)

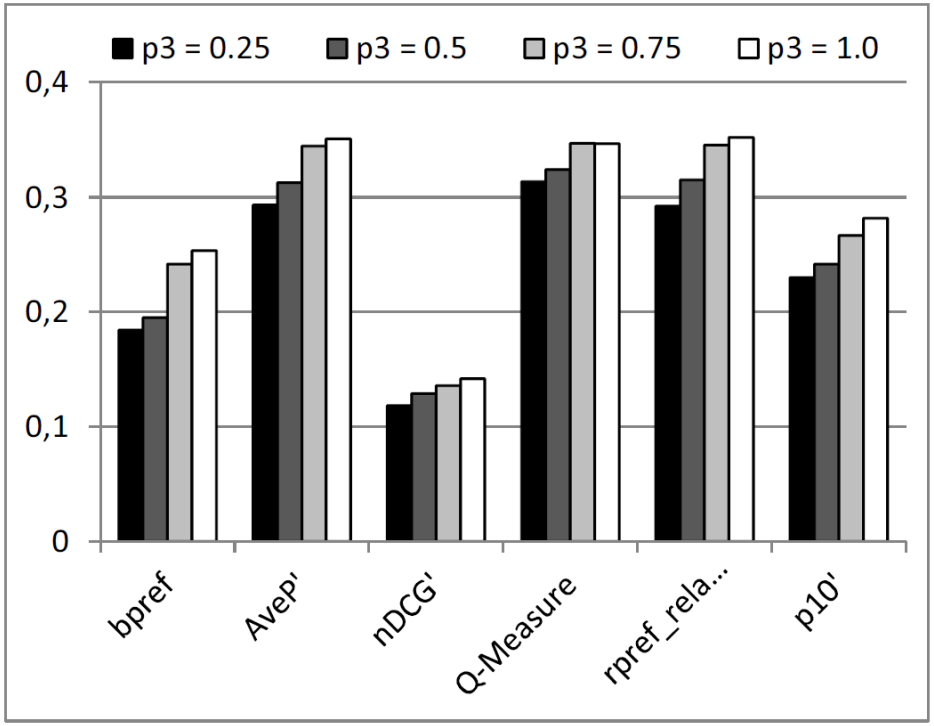


(a) bpref

- In all cases, the random lists are worse than the initial
- When the number of relevant hits in the top positions is high, re-ranking can affect negatively the results
  - Even in this case, the number of relevant-for-sure hits was increased for the majority of cases!

Effectiveness of Results Re-ranking

- Effect of Entity enrichment using DBpedia's "subject" property
  - This property provides categories/groups in which the corresponding entity belongs
- Negative effect
  - Better results when no entity enrichment (i.e., *p3 = 1.0*)
  - Worse results for larger probability to move to a property node (i.e., for small p3)



Applications

- The search system **X-Search**
- The annotation system **Theophrastus**

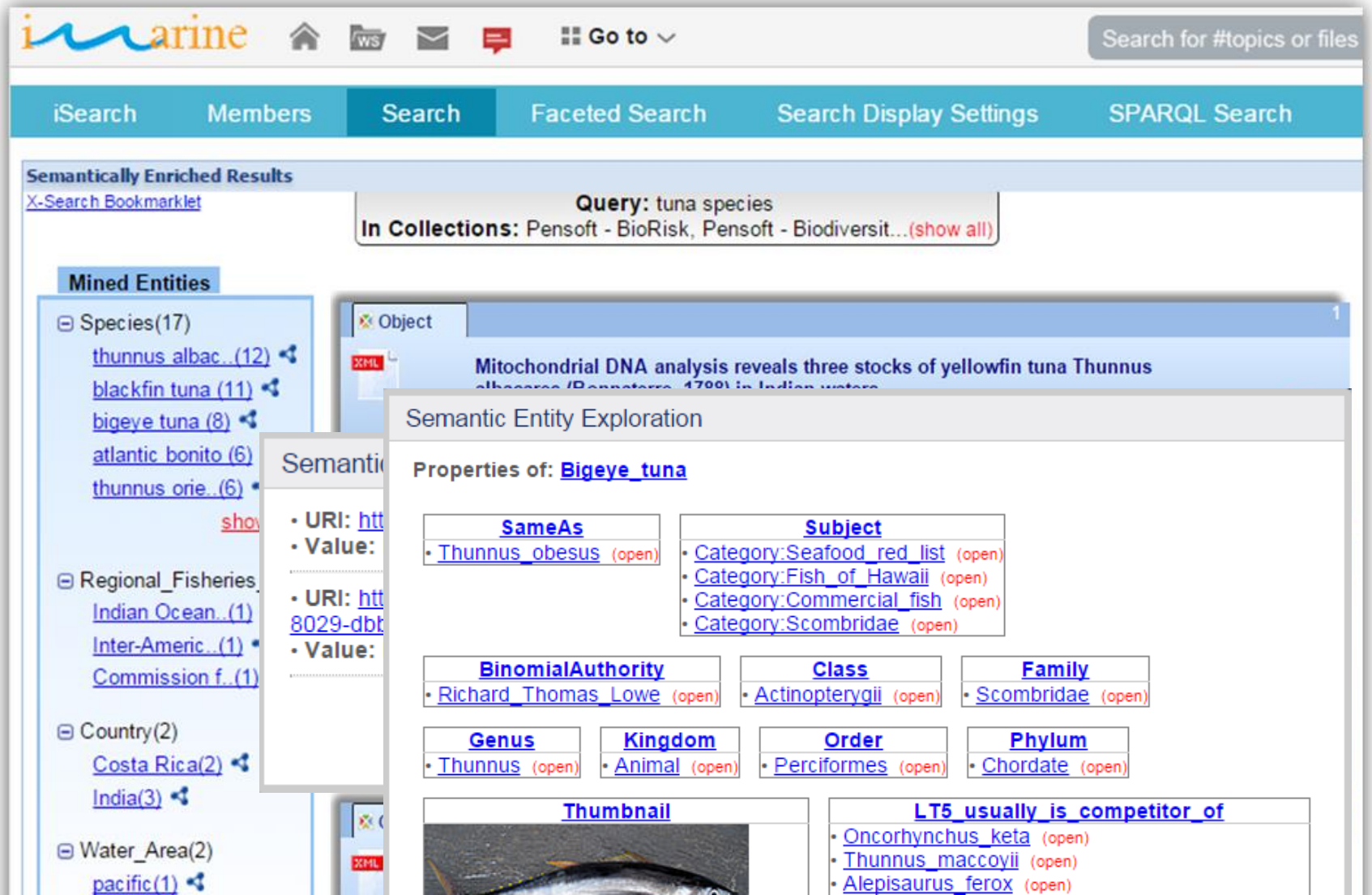
The search system "X-Search"

http://www.ics.forth.gr/isl/X-Search

- **Meta-search system** that supports the proposed semantic analysis process and configurability
- Functionality:
  - Real-time and configurable NEE using X-Link
  - Entity-based Faceted Search
  - On-demand Entity Exploration (based on the specified template queries)
- X-Search has been deployed and used in the following contexts:
  - For **Marine Search** (in the context of the **iMarine** project)
    - Deployment within an e-infrastructure serving marine biologists
  - For **Patent Search** (in the context of the **PerFedPat** project)

The search system "X-Search"

http://www.ics.forth.gr/isl/X-Search



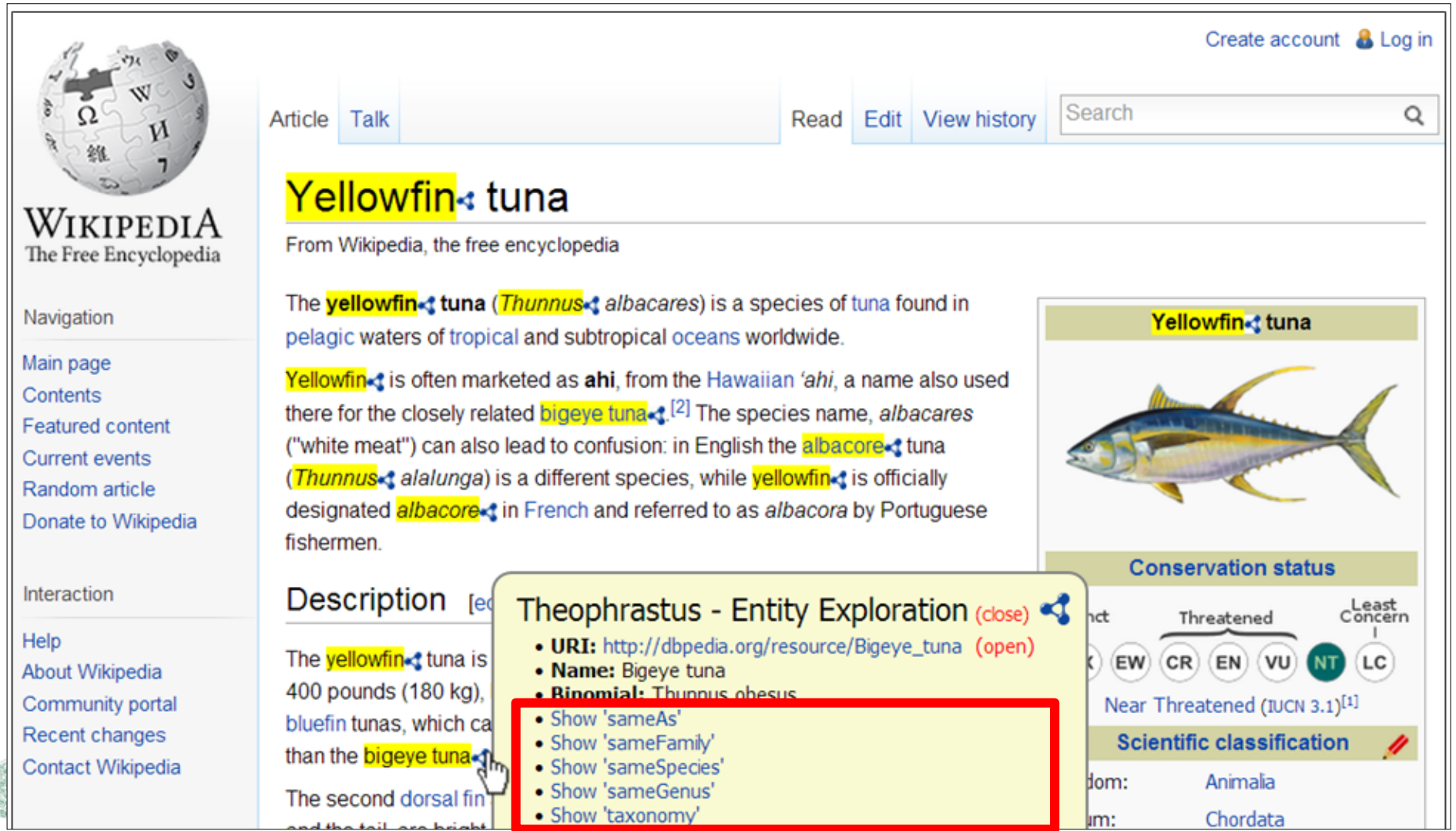
The annotation system "Theophrastus"

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

- Automatic Annotation of Web Documents
- Functionality:
  - Real-time and configurable NEE using X-Link
  - On-demand Entity Exploration based on "**Information Cards**"
  - **Information Card**
    - Piece of information regarding an entity
    - Corresponds to an Enrichment Template Query
    - E.g., for a fish species:
      - *Taxonomy*
      - *Predatory fishes*
      - *Other species belonging to the same genus*
      - …

The annotation system "Theophrastus"

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



Conclusion

- Synopsis of Contributions
- Directions for Future Research

Synopsis of Contributions

- **Integration** of Search Results with **Linked Data** for advancing Exploratory Search
  - **Entities** are the "glue"
- **General** and **Configurable** **semantic analysis** process
  - Provides a way of making the Linked Data accessible to the end-users
  - Exploitable in different contexts (*faceted search, results re-ranking, …*)
  - Can be applied at query time, no user effort

- **Focus**
  - **Configurability**: How one can **specify** the "entities of interest" and the related (useful for the search context) semantic information?
  - **Ranking of Entities and Properties**: How to **select** the semantic information (entities and properties) that better characterizes the search results?
  - **Re-Ranking of Search Results**: How to **exploit** the related semantic information (entities and properties) for producing a better ranking for the search results?

Configuring Named Entity Extraction

- Generic (Linked Data-based) Configuration Model
  - Open NEE Configuration Model
  - Open Annotation Extension
  - X-Link

- User Study
  - The majority (**80%**) of the participants managed to successfully configure the system **within a few minutes**
- Case study
  - Even querying a publicly available knowledge base (DBpedia) we can support the exploitation of LOD at real-time
- Bottleneck
  - Reliability (and performance) of existing semantic knowledge bases

Ranking of Entities and Properties

- Probabilistic (biased Random Walk-based) scheme for **ranking** entities and properties
  - Exploits the scoring/ranking of search results
  - Produces Top-K Semantic Graphs related to the search results

- Comparative evaluation with other ranking methods
  - User study in a Marine Search scenario
  - about **22%** better ranking compared to a Spreading Activation algorithm and **43%** compared to the plain (not biased) PageRank algorithm
- Efficiency of the entire semantic Analysis process
  - 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 (<4''), even if we query an online knowledge base at real-time
  - We can bound the maximum response time for being scalable

Re-Ranking of Search Results

- Random Walk model for **re-ranking** the list of retrieved results
  - Exploits the extracted entities
  - Promotes low-ranked but relevant hits containing important (for the search context) entities

- Evaluation using **TREC** datasets
  - Clinical Decision Support Track
  - Notable improvement of the list of results
    - 40% better bpref, 20% better AveP' in average
  - Negative effect when:
    - Entity enrichment is applied
    - Big number of relevant hits in the top-ranked positions

Future Research

- Stochastic Ranking
  - Consider User Actions in an interactive search context (Faceted Search)
  - Exploit the Query Terms and the Textual Data
  - Apply in other contexts (e.g., query recommendation)

- Interacting with Top-K Semantic Graphs
  - How to visualize a top-K semantic graph
  - How the top-K semantic graph should affect the interaction between user and system

- Named Entity Extraction and Configurability
  - Interoperability: model that can wholly describe the functionality, the API and the configurability of a NEE service
  - Dynamically discover and use NEE services by exploiting only standard Web protocols, without needing to set up a corresponding service

Publications (1/3)

- P. Fafalios, I. Kitsos, Y. Marketakis, C. Baldassarre, M. Salampasis and Y. Tzitzikas, "*Web Searching with EntityMining at Query Time*", Proceedings of the 5th Information Retrieval Facility Conference, IRFC 2012, Vienna, Austria, July 2012

- P. Fafalios, M. Salampasis and Y. Tzitzikas, "*Exploratory Patent Search with Faceted Search and Configurable Entity Mining*", Proceedings of the 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13), Moscow, Russia, March 2013.

- P. Fafalios and Y. Tzitzikas, "*X-ENS: Semantic Enrichment of Web Search Results at Real-Time*", Proceedings of the 36th International ACMSIGIR Conference on Research and Development in Information Retrieval, Demo Paper, SIGIR 2013, Dublin, Ireland, 28 July - 1 August 2013

- P. Fafalios, M. Baritakis and Y. Tzitzikas, "*Configuring Named Entity Extraction through Real-Time Exploitation of Linked Data*", 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14), Thessaloniki, Greece, June 2014.

Publications (2/3)

- P. Fafalios and P. Papadakos, "*Theophrastus: On Demand and Real-Time Automatic Annotation and Exploration of (Web) Documents using Open Linked Data*", Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 29 (2014), Elsevier

- 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

- P. Fafalios and Y. Tzitzikas, *"Exploratory Professional Search through Semantic Post-Analysis of Search Results"*, Professional Search in the Modern World, Lecture Notes in Computer Science (LNCS), Springer, 2014

- P. Fafalios, P. Papadakos and Y. Tzitzikas, *"Enriching Textual Search Results at Query Time using Entity Mining, Linked Data and Link Analysis"*, International Journal of Semantic Computing, Vol. 08, No. 04 (2014),World Scientific

Publications (3/3)

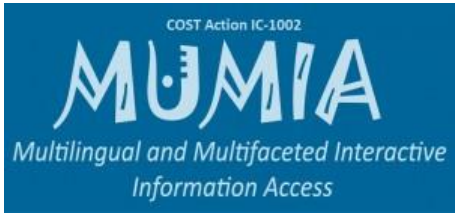
- P. Fafalios and Y. Tzitzikas, *"Type-ahead Exploratory Search through Typo and Word Order Tolerant Autocompletion"*, Journal of Web Engineering, Vol. 14, No. 1 & 2 (2015), Rinton Press

- P. Fafalios, M. Baritakis and Y. Tzitzikas, *"Exploiting Linked Data for Open and Configurable Named Entity Extraction"*, International Journal on Artificial Intelligence Tools, World Scientific, 2015

- P. Fafalios and Y. Tzitziks, *"Entity-based Stochastic Analysis of Search Results for Query Expansion and Results Re-Ranking"*, Text Retrieval Conference (TREC'15), Gaithersburg, Maryland, USA, 2015

Thank you

Questions?



Legacy
«Maria Michael Manasaki»