

# Web Searching with Entity Mining at Query Time

P. Fafalios<sup>1</sup>, I. Kitsos<sup>1</sup>, Y. Marketakis<sup>1</sup>, C. Baldassarre<sup>2</sup>,  
M. Salampanis<sup>3</sup>, Y. Tzitzikas<sup>1</sup>

1: Institute of Computer Science, FORTH-ICS, and

Computer Science Department, University of Crete, GREECE

2: Food and Agriculture Organization of the United Nations (FAO UN, Rome)

3: Vienna Univ. of Technology, Institute of Software Technology and Interactive Systems



IRF Conference, Vienna, July 2012

## The topic in one slide

**Exploratory search** services that bridge the gap between the responses of “**non semantic**” search systems (e.g. keyword search) and **semantic** information (e.g. SPARQL endpoints) through a innovative dynamic **entity**-based integration approach that is performed at **query time** (key point: no human effort !)

The provided services analyze (in various ways) and semantically enrich the returned results for satisfying recall-oriented information needs (snippet-based results clustering, entity name mining, semantic enrichment, gradual restriction of results etc. ).

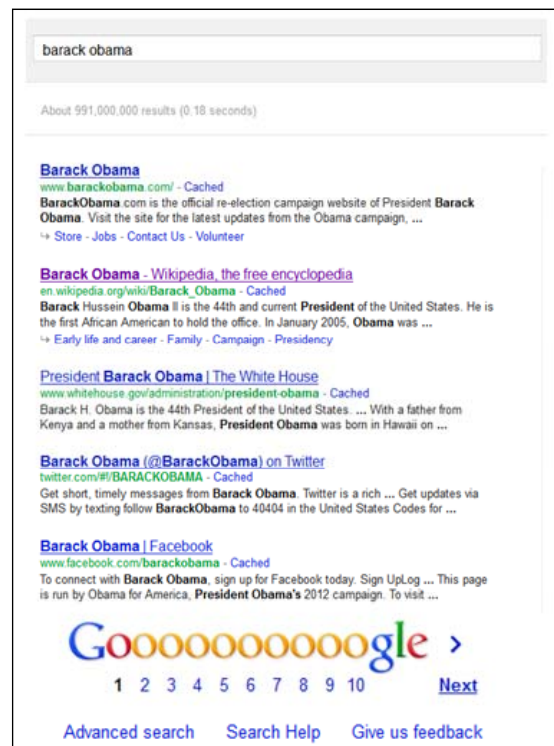
# Outline

- Motivation
- Our approach
  - The Process
  - Challenges and how we tackled them
- Experimental Results
  - Entity Ranking
  - Contents Mining vs. Snippet Mining
  - Exploiting Linked Open Data
- Case Studies
  - Fisheries and Aquaculture Publications
  - Patent Search
- Conclusion and Future Work

FORTH-ICS, IRF2012

## Motivation (1/2): Common *Web Search Engines*

- Common Web Search Engines return a ranked list of pages
- Users
  - have to explore the answer linearly
  - tend to look only at the first page of results
  - rarely exploit the available metadata (or advanced search forms)
  - adopt a trial-and-error approach



A screenshot of a Google search result for the query "barack obama". The search bar at the top shows the query and the number of results: "About 991,000,000 results (0.18 seconds)". The results are listed below, each with a title, URL, and a brief snippet. The results include:

- Barack Obama** - [www.barackobama.com/](http://www.barackobama.com/) - Cached  
BarackObama.com is the official re-election campaign website of President **Barack Obama**. Visit the site for the latest updates from the Obama campaign, ...  
→ Store - Jobs - Contact Us - Volunteer
- Barack Obama** - [Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Barack_Obama) - Cached  
en.wikipedia.org/wiki/Barack\_Obama - Cached  
**Barack Hussein Obama II** is the 44th and current **President** of the United States. He is the first African American to hold the office. In January 2005, **Obama** was ...  
→ Early life and career - Family - Campaign - Presidency
- President Barack Obama | The White House** - [www.whitehouse.gov/administration/president-obama](http://www.whitehouse.gov/administration/president-obama) - Cached  
Barack H. Obama is the 44th President of the United States. ... With a father from Kenya and a mother from Kansas, **President Obama** was born in Hawaii on ...
- Barack Obama (@BarackObama) on Twitter** - [twitter.com/#/BARACKOBAMA](https://twitter.com/BARACKOBAMA) - Cached  
Get short, timely messages from **Barack Obama**. Twitter is a rich ... Get updates via SMS by texting follow **BarackObama** to 40404 in the United States Codes for ...
- Barack Obama | Facebook** - [www.facebook.com/barackobama](http://www.facebook.com/barackobama) - Cached  
To connect with **Barack Obama**, sign up for Facebook today. Sign UpLog ... This page is run by Obama for America, **President Obama's** 2012 campaign. To visit ...

At the bottom of the screenshot, the Google logo is displayed with the word "Google" in its characteristic multi-colored font, followed by a right-pointing arrow. Below the logo are the numbers 1 through 10, and the word "Next" to the right. At the very bottom, there are three links: "Advanced search", "Search Help", and "Give us feedback".

FORTH-ICS, IRF2012

# Motivation (2/2): Emerging *Entity Search Engines*

- Existing **Entity Search Services**:
  - Provide the user with **entities** and **relationships** between these entities
  - Do not provide links to web pages
- Nevertheless:
  - Are still in their infancy (not used in many common and real world tasks)



The screenshot shows the EntityCube search interface. At the top, there's a search bar with 'barack obama' entered. Below the search bar, there are tabs for 'All Results', 'Relationship', 'Bio', 'Tag', 'Profession', 'News', 'SNS', 'Quote', 'Year', and 'Publication'. The 'All Results' tab is selected. On the left, there's a 'PEOPLE' section with a list of candidates: John McCain, Hillary Clinton, and Joe Biden. Each name is followed by a category label (e.g., 'candidates' for McCain and Clinton, 'politicians' for Biden) and a small orange circle icon. To the right of this list is a large profile card for Barack Obama, featuring a photo of him and a detailed description of his role as the 44th President of the United States, including his birth date, place, and various occupations.

FORTH-ICS, IRF2012

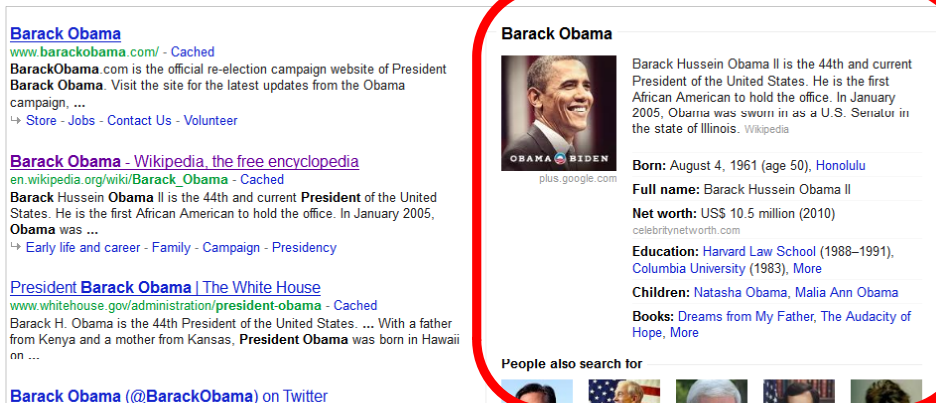
## Our Approach

Do not change the way users search for information!

**Enrich** the classical (keyword based) interaction scheme of WSEs with Named **Entity** Mining (NEM) at **query time**.

*Combine the pros of both families of systems*

Similar (but much more simple) functionality recently announced (May'12) by Google: **"Google Knowledge Graph"**



The screenshot shows a Google search result for 'Barack Obama'. On the left, there are several search results from various sources, including the official campaign website, Wikipedia, and the White House. On the right, a Knowledge Graph card is displayed, which is highlighted with a red rounded rectangle. This card includes a photo of Barack Obama, his full name, birth date and place, net worth, education, and children. Below the card, there's a section for 'People also search for' with small thumbnail images of related figures.

FORTH-ICS, IRF2012

# From an information integration point of view



- *Entity names* are used as the “glue” for automatically connecting documents with data (and knowledge).
- This approach does not require deciding or designing an integrated schema/view, nor mappings between concepts as in knowledge bases, or mappings in the form of queries as in the case of databases.
- Entities can be identified in **documents**, **data**, **database cells**, **metadata** attributes and **knowledge bases**.

FORTH-ICS, IRF2012

## Our Process

We focus on the following process

1. *Discover entities in the top hits of keyword search results*
2. *Group the entities according to their categories*
3. *Exploit the entities in faceted search-like interaction scheme (session-based)*
4. *Exploit Semantic Data (the LOD cloud) for further semantic information (also for 1)*

FORTH-ICS, IRF2012

# Prototype

The screenshot shows the 'entity mining' interface with the search term 'barack obama'. The search results are categorized into 'Person' (1427 entities) and 'Organization' (842 entities). A red box highlights the 'Person' category, and another red box highlights the 'Organization' category. A third red box highlights the 'find its entities' link. A fourth red box highlights the '100 - results to mine' and 'mine only snippets' options. A fifth red box highlights the 'show all' link. A yellow box contains the following text:

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

Results of selected entities: [reset](#)

**Person** (1427 entities)

- Barack Obama (16) ↕
- Michelle Obama (19) ↕
- George W. Bush (16) ↕
- Ann Dunham (15) ↕
- Craig Robinson (15) ↕
- Joe Biden (13) ↕
- John McCain (8) ↕
- Kennedy (9) ↕
- Sarkozy (8) ↕
- Clinton (6) ↕

**Organization** (842 entities)

- Harvard (14) ↕
- White House (18) ↕
- Congress (14) ↕
- University of Hawaii (10) ↕
- Columbia University (8) ↕

Barack Obama - Wikipedia, the free encyclopedia  
Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama ...  
[http://en.wikipedia.org/wiki/Barack\\_Obama](http://en.wikipedia.org/wiki/Barack_Obama) - find its entities

Barack Obama  
BarackObama.com is the official re-election campaign website of President Barack Obama. Visit the site for the latest updates from ...  
<https://www.facebook.com/barackobama> - find its entities

Quote of the day: "This is a good first step, but it is only a step. Congress needs to pass the rest of my American Jobs Act so that can create jobs and put ..."  
<https://www.facebook.com/barackobama> - find its entities

FORTH-ICS, IRF2012

# Prototype (cont)

The screenshot shows the 'entity mining' interface with the search term 'barack obama'. The search results are categorized into 'Person' (1427 entities) and 'Organization' (842 entities). A red box highlights the 'Person' category, and another red box highlights the 'Organization' category. A third red box highlights the 'find its entities' link. A fourth red box highlights the '100 - results to mine' and 'mine only snippets' options. A fifth red box highlights the 'show all' link. A yellow box contains the following text:

- Exploitation for restricting the focus

Results of selected entities: [reset](#)

**Person** (1427 entities)

- Barack Obama (16) ↕
- Michelle Obama (19) ↕
- George W. Bush (16) ↕
- Ann Dunham (15) ↕
- Craig Robinson (15) ↕
- Joe Biden (13) ↕
- John McCain (8) ↕
- Kennedy (9) ↕
- Sarkozy (8) ↕
- Clinton (6) ↕

**Organization** (842 entities)

- Harvard (14) ↕
- White House (22) ↕
- Congress (14) ↕

Barack Obama  
BarackObama.com is the official re-election campaign website of President Barack Obama. Visit the site for the latest updates from the Obama campaign, ...  
<http://www.barackobama.com/> - find its entities

About Barack Obama — Barack Obama  
Barack Obama is the 44th President of the United States of America. President Obama speaking. President Obama was born in Hawaii on August 4th, 1961, to a ...  
Record - Barack Obama  
my was losing more than ...  
d quickly to pass the ...  
American Recovery ...  
<http://www.barackobama.com/record> - find its entities

News for barack+obama Barack Obama - Wikipedia, the free encyclopedia  
Barack Hussein Obama II is the 44th and current President of the United

FORTH-ICS, IRF2012

# Possible Approaches for enriching Web Search and Entity Search

## 1. Off-line NEM over the entire corpus

- The corpus must be available
- We have to build an appropriate index
- The size of the index could be in the scale of the corpus!

Most existing systems

- T. Cheng et al. [2007]  
- EntityCube  
- MediaFaces

## 2. Off-line NEM over the top hits of the frequent queries\*

- We do not have to apply NEM at the entire collection
- Reduced computational effort and storage space
- Applicable also at meta-search level

WWW'12 paper

## 3. Real-time NEM over the snippets of the top hits

*default choice*

## 4. Real-time NEM over the contents of the top hits

*on demand*

The focus of this paper

\* Fafalios et al., "Scalable, Flexible and Generic Instant Overview Search", WWW' 12

FORTH-ICS, IRF2012

# Challenges & Questions

## Challenges

- Real-time response

- Perform Entity mining over the **snippets**
  - Compare the results of mining over **snippets** versus over **full contents**

- Selection/Ranking of entities

The system may discover numerous entities, but the UI has limited space

- Develop and evaluate an **entity ranking method**
  - Show only the top-10 entities. Show more entities (all) on demand.

- Exploit **Linked Open Data**

Person 1427 entities  
Barack Obama (16)  
Michelle Obama (19)  
George W. Bush (16)  
Ann Dunham (15)  
Craig Robinson (15)  
Joe Biden (10)  
John McCain (10)  
Kennedy (9)  
Sarkozy (8)  
Clinton (6)

show all

FORTH-ICS, IRF2012

# Entity Ranking – Our approach

- We focus on ranking methods that:
  - Do not rely on any log analysis (this is aligned with the dynamic nature of our approach)
- Candidate ranking methods:
  1. **Count** the documents in which the entity appears (i.e. its frequency) and take into account the **rank of the documents** that contain the entity.

$$Score_{rank}(e) = \sum_{a \in docs(e)}^{Set\ of\ returned\ hits} ((|A| + 1) - \underbrace{rank(a)}_{Its\ position\ in\ the\ answer})$$

2. Take into account the words of the entity name and the **query string**, and tolerate small differences with Edit Distance.
3. Consider **both** perspective: adopt the harmonic mean of the above scores.

FORTH-ICS, IRF2012

## On Exploiting Linked Open Data (1/2)

*“Exploiting LOD is more dynamic, affordable and feasible than an approach that require the system to store and maintain its own KB”*

### Our objectives

- Allow the user to get some basic information about an entity, without needing to submit new queries. Ability to continue browsing the related entities
- Offer a flexible configuration method. For specifying the categories and entities that are interesting for the application at hand

**Person** (1427 entities)

- Barack Obama (16) ↗
- Michelle Obama (19) ↗
- George W. Bush (16) ↗
- Ann Dunham (15) ↗
- Craig Robinson (15) ↗
- Joe Biden (10) ↗ ↗
- John McCain (10) ↗
- Kennedy (9) ↗
- Sarkozy (8) ↗
- Clinton (6) ↗



**Nicolas Sarkozy**  
Current President of France  
Birth date: 1955-01-28  
Birth place: Paris, France  
Profession: Lawyer  
Web site: <http://www.sarkozy.fr>  
Page: [http://en.wikipedia.org/wiki/Nicolas\\_Sarkozy](http://en.wikipedia.org/wiki/Nicolas_Sarkozy)

We offer this on demand: By clicking the icon the system checks if that entity lies in a LOD dataset (by performing a SPARQL query); If yes, it collects more information about that entity.

Configurability: The system allows specifying one or more appropriate LOD datasets for each category of entities. E.g. *GeoNames* for entities in “Location”, *DBpedia* for entities in “Organization”, “Person” and “Location”, etc.)

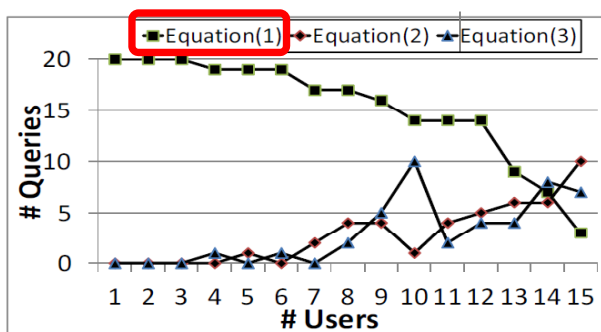
Ability to define new categories and the queries that return entities of interest.

FORTH-ICS, IRF2012

# EXPERIMENTAL RESULTS

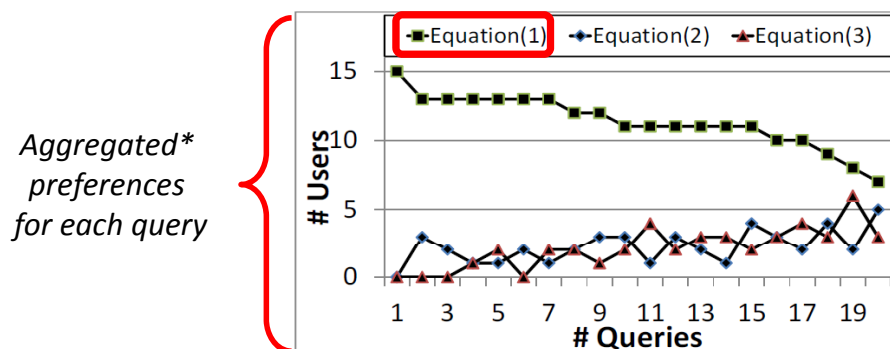
FORTH-ICS, IRF2012

## Entity Ranking – Evaluation by Users



Aggregated\* preferences for each user

- 15 users
- 20 queries
- 3 ranking methods
- For each query, user can select one, two, or all rankings.



Aggregated\* preferences for each query

\*The aggregation was based on the plurality ranking (by considering only the most preferred options)

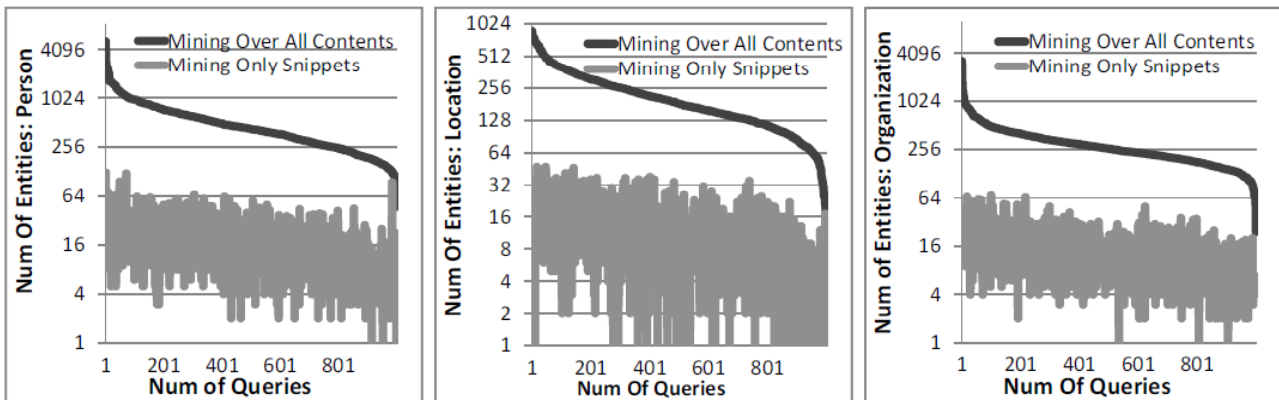
FORTH-ICS, IRF2012



# Contents Mining vs. Snippet Mining (1/3)

- Number of identified entities over all contents vs. over only snippets.

- 1000 queries
- NEM over the top-50 hits



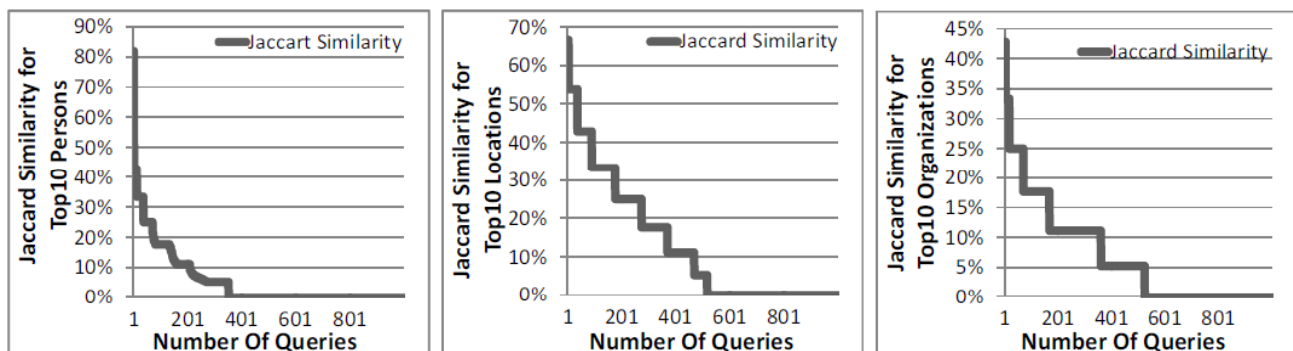
*Contents mining yields around 20 times more entities than snippet mining*

FORTH-ICS, IRF2012

# Contents Mining vs. Snippet Mining (2/3)

- Similarity of the top-10 entities

*\* Adopting the 1<sup>st</sup> Ranking approach*



*Jaccard Similarity:*

- < 30% for the majority of the queries
- 0% for about 60% of the queries

FORTH-ICS, IRF2012

# Contents Mining vs. Snippet Mining (3/3)

- Computational and Memory Costs

	Snippet Mining	Contents Mining
Entities per hit:	1.2	10.1
Overall time*:	1.5 seconds	78 seconds
Main memory footprint for one query (for the top-50 hits):	37 MB	300 MB

about 36% of total time

about 60% of total time

\* Overall time = retrieve results + **download contents** + **apply NEM at results** + apply NEM at query + create representations + rank entities

FORTH-ICS, IRF2012

## CASE STUDIES

FORTH-ICS, IRF2012

# Case Study: Fisheries and Aquaculture Publications (1/2)



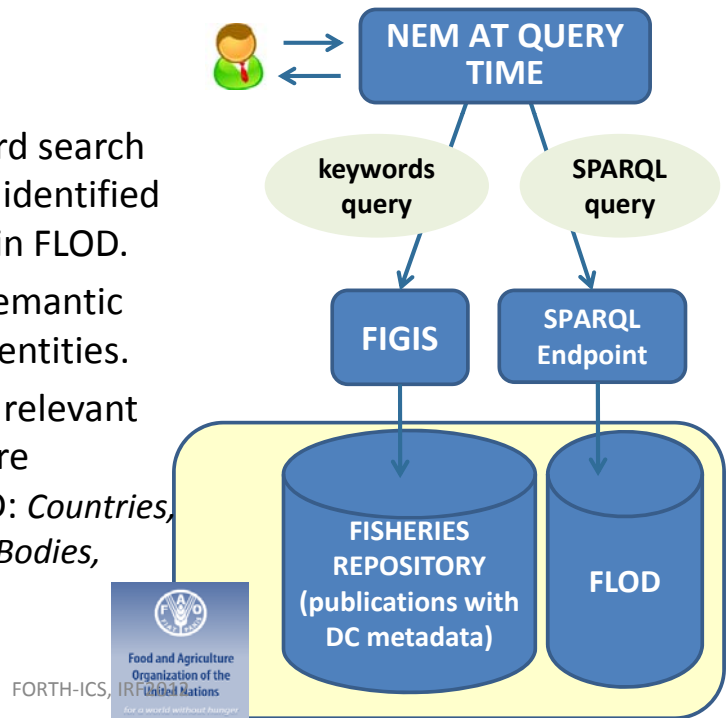
Domain:

- **FAO publications** about fisheries and aquaculture

Objective:

- Identify how to enrich keyword search with entity mining where the identified entities are linked to entities in FLOD.
- Create and serve (dynamic) semantic description for the identified entities.

We have identified the following relevant categories of entities which are semantically described in LOD: *Countries, Water Areas, Regional Fisheries Bodies, Marine Species*



FORTH-ICS, IR



# Case Study: Fisheries and Aquaculture Publications (2/2)



Identified entities linked to Entities in FLOD

mining+clustering

tuna Search

FIGIS results

Country (38 entities)

- France (8)
- Spain (5)
- Comoros (5)
- Mexico (4)
- Kenya (4)
- Madagascar (3)
- Spain (1)
- Maldives (1)
- Mauritius (1)
- Mozambique (1)

Species (26 entities)

- Adriatic trout (6)
- African red snapper (4)
- Amazon croaker (1)

Book (286931) - Report of the Special Meeting on RECOFI Consolidation and Development. Rome, Italy, 11-12 May 2011

The Special Meeting on Consolidation and Development of the Regional Commission for Fisheries (RECOFI) was held at the FAO headquarters in Rome, Italy, from 11 to 12 May 2010. The objective of the special meeting was to consider ways and means to enhance the role of RECOFI as a regional fisheries management organization. The ...

<http://www.fao.org/docrep/013/i2035e/i2035e00.pdf>

Fisheries Linked Open Data (FLOD)

- maldives

Entity label in other languages: maldives(fr), maldives(en), maldivas(es)

Semantic Description obtained by (SPARQL) querying FLOD. The user can continue browsing

FORTH-ICS, IR

## Case Study: Patent Search (professional search in general)

Entity identification and analysis could become a significant aid to professional searches and can be seen (together with other text analysis technologies) as becoming the cutting edge of information retrieval science.

### Patent Search

- Missing relevant documents is unacceptable in patent search (*recall oriented search procedure*). Retrieval of all relevant documents is usually necessary
- The related documents contain plenty of named entities
  - Companies
  - Countries
  - Persons
  - Product types
  - Laws

FORTH-ICS, IRF2012

## Related Work: Google Knowledge Graph *although announced after the submission of this paper*

The image shows a Google search result for "Barack Obama". On the left, there are several search results from various sources, including the official campaign website, Wikipedia, and The White House. On the right, a Knowledge Graph panel is highlighted with a red rounded rectangle. This panel features a portrait of Barack Obama and provides a semantic description of him, including his birth date and location, full name, net worth, education, and children. Below the main description, there is a section titled "People also search for" with small thumbnail images of related figures.

Google Knowledge Graph tries to “understand” user’s query and presents a semantic description of what user is maybe searching.

- It does **NOT locate entities** in the search results.
- It does **NOT group the results according to the discovered entities**.
- It is **NOT a recall oriented search**
- If the query is not a known entity, the user does NOT get any semantic description

FORTH-ICS, IRF2012

## Concluding Remarks

- Enhancing web searching with Named Entity Mining (*at query time, without any preprocessing*):
  - Gives the user an overview of the answer space
  - Allows the user to restrict his focus on a part of the answer
  - Is convenient for user needs that require collecting entities
- 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)
- By exploiting LOD for enriching the semantic descriptions of identified entities
  - the user gets useful information about one entity without submitting new queries
  - the user can start browsing the entities that are linked to that entity.

FORTH-ICS, IRF2012

## Issues for Further Research

- For the case of *contents mining*: take into account the *local score* of an entity in the document at entity ranking
- Comparatively evaluate with users the top-10 entities from snippets mining with the top-10 from contents mining.
- Long term vision:
  - Mine not only correct entities but probably entire conceptual models (and entity relationships)
  - Support the interaction paradigm of faceted search over such (crispy or fuzzy) semantic models

FORTH-ICS, IRF2012

*Thank you for your attention*

Demos:

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

FORTH-ICS, IRF2012