Semantic Integration of Marine Data

Yannis Tzitzikas
Computer Science Department, University of Crete, GREECE &
Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)
tzitzik@ics.forth.gr

BlueBRIDGE Webinar
28 June 2017
Outline

• Motivation (for semantic integration)

• Difficulties and Requirements

• Case Study: Marine Data
  • process, techniques, outcomes, exploitation

• Conclusion

Time plan: 40’ presentation, 15’ questions and discussion
Huge amounts of data are available and this amount constantly increases. Almost everyone produces data (and everything will produce data). Almost everyone needs data (and everything will need data).
Motivation

However data and information are not integrated.
Motivation

However data and information are **not integrated**.

Hundreds or thousands of CKAN catalogs each containing hundreds or thousands of datasets.
However, in several domains and applications one has to fetch and assemble pieces of information coming from more than one datasets/sources for being able to answer complex queries (that are not answerable by individual sources) or analyze the integrated data.

This is of paramount importance for e-science especially for large-scale scientific questions such as global warming, invasive species spread, and resource depletion and concerns several domains:

- Biodiversity domain
- Cultural Domain
- E-Government
- Science in general
- ...

- Personal data
It has been written that

- Data scientists spend from 50 percent to 80 percent of their time in collecting and preparing unruly digital data, before it can be explored for useful nuggets.
- If you’re trying to reconcile a lot of sources of data that you don’t control it can take 80% of your time
- One-Third of BI Pros Spend Up to 90% of Time Cleaning Data
Indicative Complex Queries

Marine Domain

• Given the scientific name of a species (say *Thunnus Albacares*), find the ecosystems, waterareas and countries that this species is native to, and the common names that are used for this species in each of the countries, as well as their commercial codes.

Cultural Domain

• Give me all paintings of El Greco that are now exhibited in Greece and their location, as well as all articles or books about these paintings between 2000 and 2016.
• Give me the paintings of El Greco referring to persons that were born between 0 and 300 BC.
• Give me all events related to El Greco that will take place this month in Heraklion.
Why integration is difficult?

- Datasets are kept or produced by different organizations in different formats, models, locations, systems.

- The same real world entities or relationships are referred with different names and in different natural languages (natural languages have synonyms and homonyms)

- Datasets usually contain complementary information

- Datasets can contain erroneous or contradicting information

- Datasets about the same domain may follow different conceptualizations of the domain
348 common names in 82 different languages!
Why integration is difficult?

... names

Argentina

Country in South America

Greater Argentine (*Argentina silus*)

Argentine (*Argentina sphyraena*)
Why integration is difficult?

... complementary views
Why integration is difficult?

...different conceptualizations

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
General **Requirements** or Tasks related to Information Integration

- Dataset **discovery**
- Dataset **selection** (or sub-dataset selection)
- Dataset **access** and query
- Fetch and **transformation** of data
- Data and dataset **linking**
- Data **cleaning**
- Data **completion** (through context, inference, prediction or other methods)
- Management of data **provenance**
- Measuring and testing the **quality** of datasets (especially the integrated)
- Management (and understanding) of the **evolution** of datasets
- **Monitoring**, production of **overviews**, visualization of datasets
- Interactive browsing and **exploration** of datasets
- Data **summarization**, **analysis**, **preservation**

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
In general there are two main approaches for integration:

**The Warehouse approach** (materialized integration)
- S1
- S2
- S3
- Extract, Transform, store
- Warehouse
- Query
- Answer
- Mappings

**The Mediator approach** (virtual integration)
- S1
- S2
- S3
- Mediator
- Query translation
- Mappings
- Query
- Answer
Main approaches for integration (cont.)

The **Warehouse** approach  
(*materialized integration*)

**Benefits**
- **Flexibility in transformation logic** (including ability to curate and fix problems)
- **Faster responses** (in query answering but also in other tasks, e.g. if one wants to use it for applying an entity matching technique).
- **Decoupling of access load** from the underlying sources.
- **Decoupling of the release management of the integrated resource** from the management cycles of the underlying sources.

**Shortcomings**
- You have to pay the cost for hosting the warehouse.
- You have to refresh periodically the warehouse.

The **Mediator** approach  
(*virtual integration*)

**Benefits**:
- One advantage (but in some cases disadvantage) of **virtual integration** is the real-time reflection of source updates in integrated access.

**Comment**:
- The higher complexity of the system (and the quality of service demands on the sources) is only justified if immediate access to updates is indeed required.

---

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
Case Study: Marine Species
Context: iMarine

**Id:** An FP7 Research Infrastructure Project (2011-2014)

**Final goal:** launch an initiative aimed at establishing and operating an e-infrastructure supporting the principles of the Ecosystem Approach to fisheries management and conservation of marine living resources.

**Partners:**

- Consiglio Nazionale delle Ricerche
- FORTH Institute of Computer Science
- FAO
- IRD
- HELLENIC REPUBLIC National and Kapodistrian University of Athens
- CERN
- FIN
- UNESCO
- OBIS
- ENGINEERING
terradue 2.0
ERCIM
Continuation in BlueBRIDGE

BlueBRIDGE (Building Research environments for fostering Innovation, Decision making, Governance and Education to support Blue growth), H2020-EINFRA-2015-1

Sept. 2015- Feb. 2018

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
Marine Species Information: in several sources

WoRMS: World Register of Marine Species
Registers more than 200K species

ECOSCOPE- A Knowledge Base About Marine Ecosystems (IRD, France)

FLOD (Fisheries Linked Data) of Food and Agriculture Organization (FAO) of the United Nations

FishBase: Probably the largest and most extensively accessed online database of fish species.

DBpedia
Marine Information:

- **Taxonomic information**
- **Ecosystem information** (e.g. which fish eats which fish)
- **Commercial codes**
- **General information, occurrence data**, including information from other sources
- **General information, figures**
Marine Information:

- Web services (SOAP/WSDL)
- RDF + OWL files
- SPARQL Endpoint
- Relational Database
- SPARQL Endpoint

Accessed through different technologies.
How to integrate?
Scope Control (how to control it?)

• We use the notion of competency queries.
  • A competency query is a query that is useful for the community at hand, e.g. for a human member, or for building applications for that domain

• Indicative competency queries for our running example:

<table>
<thead>
<tr>
<th>#Query</th>
<th>For a scientific name of a species (e.g. Thunnus Albacares or Poromitra Crassiceps), find/give me:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁</td>
<td>the biological environments (e.g. ecosystems) in which the species has been introduced and more general descriptive information of it (such as the country)</td>
</tr>
<tr>
<td>Q₂</td>
<td>its common names and their complementary info (e.g. languages and countries where they are used)</td>
</tr>
<tr>
<td>Q₃</td>
<td>the water areas and their FAO codes in which the species is native</td>
</tr>
<tr>
<td>Q₄</td>
<td>the countries in which the species lives</td>
</tr>
<tr>
<td>Q₅</td>
<td>the water areas and the FAO portioning code associated with a country</td>
</tr>
<tr>
<td>Q₆</td>
<td>the presentation w.r.t Country, Ecosystem, Water Area and Exclusive Economical Zone (of the water area)</td>
</tr>
<tr>
<td>Q₇</td>
<td>the projection w.r.t. Ecosystem and Competitor, providing for each competitor the identification information (e.g. several codes provided by different organizations)</td>
</tr>
<tr>
<td>Q₈</td>
<td>a map w.r.t. Country and Predator, providing for each predator both the identification information and the biological classification</td>
</tr>
<tr>
<td>Q₉</td>
<td>who discovered it, in which year, the biological classification, the identification information, the common names - providing for each common name the language and the countries where it is used in.</td>
</tr>
</tbody>
</table>
Materialization or Mediation?

In both cases we need a unified model/schema
MarineTLO aims at being a global core model that

– provides a common, agreed-upon and understanding of the concepts and relationships holding in the marine domain to enable knowledge sharing, information exchanging and integration between heterogeneous sources

– covers with suitable abstractions the marine domain to enable the most fundamental queries, can be extended to any level of detail on demand, and

– allows data originating from distinct sources to be adequately mapped and integrated

• MarineTLO is not supposed to be the single ontology covering the entirety of what exists

Benefits:

• reduced effort for constructing mappings: this approach avoids the inevitable combinatorial explosion and complexities that results from pair-wise mappings between individual metadata formats and/or ontologies

• reduced effort for improving and evolving: the focus is given on one model, rather than many (the results are beneficial for the entire community)
MarineTLO: Query capabilities

It should allow formulating the competency queries.
Indicative examples of queries that can be formulated

1. Given the scientific name of a species, find its predators with the related taxon-rank classification and with the different codes that the organizations use to refer to them.

2. Given the scientific name of a species, find the ecosystems, waterareas and countries that this species is native to, and the common names that are used for this species in each of the countries.

The MarineTLO (version 5.0) currently contains around

• 115 classes
• 85 properties.

More in www.ics.forth.gr/isl/MarineTLO
We will focus on the **materialization case**

- i.e. on the construction and maintenance of a MarineTLO-based semantic warehouse
What process to follow?
How to connect the fetched pieces of information?

The **Semantic Approach**

- **Use URIs instead of strings**
  - You can establish links in this way
  - You can avoid the problem of homonyms
- **Use owl:sameAs to connect equivalent URIs**
  - Support of synonyms and equivalent identifiers
- **Various other semantic relationships** (useful for query interoperability)

**Linked Data** is a method of publishing structured data so that it can be interlinked and become more useful through **semantic queries**. It builds upon standard Web technologies such as **HTTP**, **RDF** and **URIs**. This enables data from different sources to be connected and queried.
Define **requirements** in terms of **competency queries**

Fetch the data from the selected sources (SPARQL endpoints, services, etc)

Transform and Ingest to the Warehouse

Inspect the **connectivity** of the Warehouse

Formulate rules creating **sameAs** relationships

Apply the rules to the warehouse

Ingest the **sameAs** relationships to the warehouse

Test and evaluate the Warehouse (using the competency queries and the conn. metrics)

Expressed over MarineTLO

MatWare, X3ML

MatWare

assembla Silk Link

MatWare

MatWare
How to link

- We need **Entity Matching**
  - Both *automatic* methods and *handcrafted rules* are required
Example: **Suffix**-based URI equivalence

\[
\text{thunnusalbacares} \equiv \text{thunnus_albacares}
\]

- **lower case conversion**
- **underscore removal**
- **prefix removal**

http://www.dbpedia.com/Thunnus_Albacares

\[
\equiv \text{http://www.ecoscope.com/thunnus_albacares}
\]

\[
\text{http://www.dbpedia.com/Thunnus_Albacares} = \text{http://www.ecoscope.com/thunnus_albacares}
\]

---

last(u): is the string obtained by (a) getting the substring after the last "/" or "/#", and turning the letters of the picked substring to lowercase and deleting the underscore letters that might exist.
Matching Rule:
If an Ecoscope individual's preflabel in lower case is the same with the attribute label of a FLOD individual then these two individuals are the same.

Example:
Entity Matching-based URI equiv.

Thunnus Albacares

thunnus albacares

http://www.fao.org/figis/flod/entities/codedentity/636cdcea-c411-43ad-97ff-00c9304f5e60

http://www.ecoscope.com/thunnus_albacares

skos:preflabel

sameAs
How to measure the quality of the warehouse?
Why it is useful to measure Connectivity?

- For assessing how much the aggregated content is connected
- For getting an overview of the warehouse
- For quantifying the value of the warehouse (query capabilities)
  - Poor connectivity affects negatively the query capabilities of the warehouse.
- For making easier its monitoring after reconstruction
- For measuring the contribution of each source to the warehouse, and hence deciding which sources to keep or exclude (there are already hundreds of SPARQL endpoints). Identification of redundant or unconnected sources

- In general **Connectivity** has two main aspects: Schema and Instance.
  - Regarding Schema Connectivity our running example uses a top level ontology (MarineTLO) and schema mappings in order to associate the fetched data with the schema of the top level ontology.
  - As regards Instance Connectivity one has to inspect and test the connectivity of the “draft” warehouse through the competency queries, and a number of connectivity metrics that we have defined and then formulate rules for instance matching.
### Connectivity Metrics Definition

<table>
<thead>
<tr>
<th>Metric Name</th>
<th>Metric Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common URIs between two Sources $S_i$ and $S_j$</td>
<td>$</td>
</tr>
<tr>
<td>Percentage of Common URIs between two Sources $S_i$ and $S_j$</td>
<td>$cur_{i,j} = \frac{</td>
</tr>
<tr>
<td>Common Literals between two Sources $S_i$ and $S_j$</td>
<td>$</td>
</tr>
<tr>
<td>Percentage of Common Literals between two Sources $S_i$ and $S_j$</td>
<td>$clit_{i,j} = \frac{</td>
</tr>
<tr>
<td>Increase in the average degree</td>
<td>$\frac{deg_W(E) - deg_S(E)}{deg_S(E)}$</td>
</tr>
<tr>
<td>Unique Triples of a Source $S_i$</td>
<td>$triples_{Unique}(S_i) = triples(S_i) \setminus (\bigcup_{1 \leq j \leq k, j \neq i} triples(S_j))$</td>
</tr>
<tr>
<td>Percentage of Unique Triples of a Source $S_i$</td>
<td>$\frac{</td>
</tr>
<tr>
<td>Complementarity factor for an entity $e$</td>
<td>$cf(e) =</td>
</tr>
</tbody>
</table>
Connectivity Metrics: 

Increase in the average Degree

<table>
<thead>
<tr>
<th>Source</th>
<th>avg degrees in sources</th>
<th>avg degrees in the warehouse</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOD</td>
<td>7.18</td>
<td>54.31</td>
<td>656.51%</td>
</tr>
<tr>
<td>WoRMS</td>
<td>3.3</td>
<td>9.93</td>
<td>201.36%</td>
</tr>
<tr>
<td>Ecoscope</td>
<td>22.84</td>
<td>165.24</td>
<td>623.6%</td>
</tr>
<tr>
<td>DBpedia</td>
<td>41.41</td>
<td>84.2</td>
<td>103.36%</td>
</tr>
<tr>
<td>FishBase</td>
<td>18.86</td>
<td>50.6</td>
<td>168.32%</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td><strong>18.72</strong></td>
<td><strong>72.86</strong></td>
<td><strong>289.21%</strong></td>
</tr>
</tbody>
</table>

The average degree, of all sources is significantly bigger than before.

- **Suffix canonicalization**
- **Entity Matching**

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
What about the *provenance*?
It is important to keep the provenance of each data in the warehouse. We have realized that the following 4 levels of provenance support are usually required:

[a] Conceptual level
[b] URIs and Values level
[c] Triple Level
[d] Query level

Level [a] can be supported by the conceptual model level. In our application context we use the MarineTLO and the transformation rules do the required transformations. Matware offers support also for levels [b]-[d]
a) **Conceptual modeling level**

Example: *Assignment of identifiers to species*

- MarineTLO models the provenance of species names, codes etc, and the Transformation rules of MatWare transform the ingested data according to this model.

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
b) URIs and Literals:

i. Adopting the namespace mechanism for URIs:
   - The prefix of the URI provides information about the origin of the data.
   - e.g. www.fishbase.org/entity/ecosystem#mediterranean_sea

ii. Ability to attach @Source to every literal coming from a Source:
   - e.g. select scientific name and authorship of Yellow Fin Tunna

<table>
<thead>
<tr>
<th>scientificName</th>
<th>year</th>
<th>authorship</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Thunnus albacares&quot;@dbpedia</td>
<td>&quot;1788&quot;@dbpedia</td>
<td>&quot;Bonnaterre&quot;@dbpedia</td>
</tr>
<tr>
<td>&quot;Thunnus albacares&quot;@worms</td>
<td>&quot;1788&quot;@worms</td>
<td>&quot;Bonnaterre&quot;@worms</td>
</tr>
</tbody>
</table>

- This policy allows formulating source-centric queries in a relative simple way:
  SELECT ?speciesname
  WHERE {
    ?species tlo:has_scientific_name ?scientificname
    FILTER(langMatches(lang(?scientificname), "worms"))
  }
c) Triple Level Provenance

- Store the fetched triples in a separate graphspace:
  - FISHBASE: http://www.ics.forth.gr/isl/Fishbase
  - DBpedia: http://www.ics.forth.gr/isl/DBpedia
  - FLOD: http://www.ics.forth.gr/isl/FLOD
  - Ecoscope: http://www.ics.forth.gr/isl/Ecoscope
  - WoRMS: http://www.ics.forth.gr/isl/WoRMS

- By asking for the graph that each triple is coming from we retrieve the provenance of the data.
- This enables refreshing only one part of the warehouse
d) Query Level Provenance

- Offer a query rewriting functionality that exploits the contents of the graphspaces for returning the sources that contributed to the query results (including those that contributed to the intermediate steps).
Example of Query Level Provenance:

**Query:** For a scientific name of a species (e.g. *Thunnus Albacares*) find the FAO codes of the waterareas in which the species is native.

**Query in SPARQL:**

```sparql
select ?faocode ?source1 ?source2
where {
  graph ?source1 {
    ecoscope:thunnus_albacares MarineTLO:isNativeAt ?waterarea .
  }.
  graph ?source2 {
    ?waterarea MarineTLO:LXrelatedIdentifierAssignment ?faocode
  }
}
```

**RESULT:**

<table>
<thead>
<tr>
<th>faocode</th>
<th>source1</th>
<th>source2</th>
</tr>
</thead>
</table>
Architecture of MatWare

- **Actions in order to create a Warehouse from scratch** one should specify
  - the type of the repository
  - the names of the graphs that correspond to the different sources
  - URL, username and password in order to connect to the repository
- **Actions in order to add a new source**
  - (a) include the fetcher class for the specific source as plug in
  - (b) provide the mapping files (schema mappings)
  - (c) include the transformer class for the specific source as a plug in
  - (d) provide the SILK rules as xml files
Integrated information about Thunnus albacares from different sources
### The resulted **MarineTLO-based Warehouse** (1/2)

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Ecoscope</th>
<th>FLOD</th>
<th>WoRMS</th>
<th>DBpedia</th>
<th>Fishbase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Scientific Names</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Authorships</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Common Names</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Predators</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Ecosystems</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Countries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Water Areas</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Vessels</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Gears</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>EEZ</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Evolution over Time

Figure 6: Triples of each version.

Figure 7: URIs and Literals.

Figure 10: Common URIs %

Figure 11: Common Literals %

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
**GRSF (Global Record of Stocks and Fisheries)**

**Fishery** is an activity leading to the harvesting of fish, within the boundaries of a defined water area. It gathers indication of human fishing activity with economic, management, biological, environmental and technological viewpoints.

**Stock** is a group of individuals of a species occupying a well-defined spatial range independent of other stocks of the same species.
1. Fetch
2. Transform
3. Dissect
4. Merge
5. Publish (for curation)
6. Curate & Validate
7. Publish (for exploitation)
8. Publish (for curation)
9. Exploit
The current contents of the GRSF KB

- 2,190 Stocks / Assessment Units
- 7,712 Fishery Activities/Descriptions
- 1204 Species
- 1181 Areas
- 97 Gear types
- 163 FlagStates
- 110 Assessment Methods
- 506 Scientific Advices
What about the *exploitation*?
Exploitation of Semantic Warehouses

A) Semantic Processing of Search Results
B) Fact Sheet Generator (web application)
C) Species Identification Tool
D) Exploitation of GRSF
E) Interactive 3D visualization
A) For **Semantic Processing of Search Results**:

The process

- **Semantic Analysis**
  - Grouping, Ranking
  - Retrieving more properties

- **Entity Mining**

- **Visualization/Interaction**
  - Faceted search, entity exploration, annotation, top-k graphs, etc.

- **OpenSearch** (web browsing)

- **The LOD cloud**

- **MarineTLO Warehouse**

- **Entities / Contents**

- **Query terms**

- **Contents**
A) For Semantic Processing of Search Results: Example XSearch

The Warehouse is used

Result of Entity Mining

Semantic Entity Exploration
- **URL:** http://www.fao.org/figis/floc/entities/codedentity/3e6d2db-1f06-4390-ac4a-9d3c8b95b5f (open)
- **Value:** yellowtail amberjack

Semantic Entity Exploration

**Properties of Yellowtail_amberjack**
- **Type:** Animal
- **SameAs:** Seriola lalandi
- **Subject:** Category: Fish of the Red Sea, Category: Fish of the Indian Ocean, Category: Seriola
- **Class:** Actinopterygii
- **Family:** Carangidae
- **Genus:** Seriola
- **Kingdom:** Animal
- **Order:** Perciformes
- **Phylum:** Chordata

The Warehouse is used

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
Example of the EntityCard of Thunnus Albacares

<table>
<thead>
<tr>
<th>Binomial Authority</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pierre_Joseph_Bonnaire (open)</td>
<td>Actinoptygii (open)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>Genus</th>
<th>Kingdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scombridae (open)</td>
<td>Thunnus (open)</td>
<td>Animal (open)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Order</th>
<th>Phylum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perciformes (open)</td>
<td>Chordate (open)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FAO Distribution</th>
<th>Is Classified By Code</th>
<th>ScientificName</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species_dist.22559 (open)</td>
<td>Yft (open)</td>
<td>Thunnus Albacares</td>
</tr>
<tr>
<td>Species_dist.22560 (open)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Has Related Indicator</th>
<th>Is Predator Of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel_density (open)</td>
<td>Sarda_orientalis (open)</td>
</tr>
<tr>
<td>I3_SpeciesYearByGearMonth_YFT_1999 (open)</td>
<td>Sthenoteuthis_Oualiensi (open)</td>
</tr>
<tr>
<td>I3_SpeciesYearByGearMonth_YFT_2001 (open)</td>
<td>Mastigoteuthis_sp (open)</td>
</tr>
<tr>
<td>I2_SpeciesByGear_YFT (open)</td>
<td>Scomberesox_saurus (open)</td>
</tr>
<tr>
<td>I3_SpeciesYearByGearMonth_YFT_2000 (open)</td>
<td>Neanchisquilla_tuberculata (open)</td>
</tr>
<tr>
<td>I6_SpeciesMap_YFT_2001_2002 (open)</td>
<td>Vinciguerra_nimbaria (open)</td>
</tr>
<tr>
<td>II_SpeciesByOcean_YFT (open)</td>
<td>Odontodactylus_scyllarus (open)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Depiction</th>
<th>Related Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>IndicatorThunnusAlbacares2 (open)</td>
<td>LayerDbStomacYFT (open)</td>
</tr>
<tr>
<td>IndicatorThunnusAlbacares1 (open)</td>
<td>LayerDbSardaraYFT (open)</td>
</tr>
<tr>
<td>IndicatorThunnusAlbacares4 (open)</td>
<td></td>
</tr>
<tr>
<td>IndicatorThunnusAlbacares3 (open)</td>
<td></td>
</tr>
<tr>
<td>PictureToBeTagged13_1 (open)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Has Taxon Id</th>
<th>Belongs To</th>
</tr>
</thead>
<tbody>
<tr>
<td>WoRMS:127027 (open)</td>
<td>Thunnus (open)</td>
</tr>
</tbody>
</table>

The Warehouse is used
A’) XSearch as a bookmarklet

Dynamic annotating of entities over any Web page

The Warehouse is used

Thunnus

From Wikipedia, the free encyclopedia

Main article: Tuna

Thunnus is a genus of ocean-dwelling ray-finned bony fish from the Scombridae (Mackerel) family. More specifically, Thunnus is one of five genera which comprise the Thunnini tribe—a tribe that is collectively (and famously) known as the tunas. Also called the true tunas or real tunas, Thunnus consists of eight species of tuna (more than half of the overall tribe), divided into two subgenera.

The word Thunnus is the Middle Latin form of the Ancient Greek: θῦνος (θῦνος) “tunny fish” — which is in turn derived from θύω (θύω), “to rush; to dart”. The first written use of the word was by Homer.

Their coloring, metallic blue on top and shimmering silver-white on the bottom, helps camouflage them from above and below. They can grow to 16 feet long and weigh over 1,000 pounds, and can swim up to 50 miles per hour when pursuing prey. Atlantic bluefin tunas are warm-blooded, which is a rare trait among fish, and are comfortable in the cold waters. Bluefin are found in Newfoundland and Iceland, as well as the tropical waters of the Gulf of Mexico and the Mediterranean Sea, where they go each year to spawn.

Due to overfishing the genus range has been significantly reduced, being effectively removed from the Black Sea, for example.

Taxonomy

This genus has eight species in two subgenera:

- Subgenus Thunnus (Thunnus):
  - Albacore, Thunnus alalunga (Bonaparte, 1860).
  - Southern bluefin tuna, Thunnus maccoyii (Castelnau, 1872).
  - Bigeye tuna, Thunnus obesus (Bonnaterre, 1788).
- Subgenus Thunnus
  - Yellowfin tuna, Thunnus albacares (Cuvier, 1829).
  - Pacific bluefin tuna, Thunnus orientalis (Temminck & Schlegel, 1840).
- Subgenus Thunnus
  - Blackfin tuna, Thunnus atlanticus (Cuvier, 1829).
- Subgenus Thunnus
  - Longtail tuna, Thunnus obesus (Cuvier, 1829).

Semantic Entity Enrichment:

- URL: http://www.ecoscope.org/ontologies/ecosystem?thunnus_obesus (open)
- Label: Bigeye tuna

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
B) Fact Sheet Generator & Android Application

Fact Sheet Generator

Ichthys

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
C) Species Identification Tool

General Pattern

- **Semantic views** can be defined over the Semantic Warehouse for selecting those parts of the SW that are useful for carrying out one particular task.

Example

- For the task of **species identification** we have defined a semantic view and then use **Preference-enriched Faceted Search** for offering to end users an interaction model appropriate for exploration and decision making.
C) Species Identification Tool

Species identification through **Preference-enriched Faceted Search** over the semantic descriptions of fish species (extracted by querying the semantic warehouse).
D) Exploitation of GRSF

- GRSF contents are exposed through the D4Science Resource Catalogue
  - Users can search and browse GRSF

**GRSF Catalogue**

**GRSF Records metadata**

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>GRSF Publisher</td>
</tr>
<tr>
<td>Version</td>
<td>1</td>
</tr>
<tr>
<td>State</td>
<td>active</td>
</tr>
<tr>
<td>Last Updated</td>
<td>February 7, 2017, 10:22 PM (UTC+02:00)</td>
</tr>
<tr>
<td>Created</td>
<td>December 15, 2016, 10:03 PM (UTC+02:00)</td>
</tr>
<tr>
<td>Exploiting fishery</td>
<td><a href="http://data.d4science.org/uri-resolver/catalogue/GRSF_Admin/product/5c7bceea-03a5-3ff6-ad51-08831653bab0">http://data.d4science.org/uri-resolver/catalogue/GRSF_Admin/product/5c7bceea-03a5-3ff6-ad51-08831653bab0</a></td>
</tr>
<tr>
<td>Product URL</td>
<td><a href="http://data.d4science.org/uri-resolver/catalogue/GRSF_Admin/product/6db361ea-ba8f-39b-a-b1ae-81b2511386d4">http://data.d4science.org/uri-resolver/catalogue/GRSF_Admin/product/6db361ea-ba8f-39b-a-b1ae-81b2511386d4</a></td>
</tr>
<tr>
<td>Product type</td>
<td>Stock</td>
</tr>
<tr>
<td>Record type</td>
<td>Aggregated</td>
</tr>
<tr>
<td>Short Title</td>
<td>penaeus esculentus FAO 71</td>
</tr>
<tr>
<td>Species</td>
<td>PRB</td>
</tr>
<tr>
<td>Status</td>
<td>Pending</td>
</tr>
<tr>
<td>Stock Id</td>
<td>prb-fao 71</td>
</tr>
<tr>
<td>Stock Name</td>
<td>Brown tiger prawn - Eastern Australia</td>
</tr>
</tbody>
</table>

**GRSF timeseries (e.g. exploitation rate)**
E) Interactive 3D Visualization of Datasets

The metrics are exploited for producing interactive 3D visualization of datasets (for providing informative **overviews** and for aiding **monitoring** the semantic warehouses over time)
This approach is general

Integrated information about Thunnus albacares from different sources

Datasets about Crete

Datasets about Toledo

Datasets about Art
A few words about the earlier steps of the process

- Dataset discovery
- Dataset selection (or sub-dataset selection)
- Dataset access and query
- Fetch and transformation of data
- Data and dataset linking
- Data cleaning
- Data completion (through context, inference, prediction or other methods)
- Management of data provenance
- Measuring and testing the quality of datasets (especially the integrated)
- Management (and understanding) of the evolution of datasets
- Monitoring, production of overviews, visualization of datasets
- Interactive browsing and exploration of datasets
- Data summarization, analysis, preservation
Objective

- Currently, in BlueBridge we are in the process of designing services that facilitate the discovery of datasets by enabling complex query answering. The approach is based on semantical enrichment of Data Catalogue Facilities.

Approach

- Enhancement of the metadata of the data catalog entries through their automatic linking with other semantic resources (i.e. DBpedia, GRSF, etc.) and provision of search features that exploit the semantics of the metadata of the published datasets.

Benefits

- Facilitation of the discovery of datasets by enabling complex query answering, e.g. *search for datasets containing information about Sparus Aurata Species, published by an iMarine organization within 2016 and 2017*. 

The big picture (core concepts and relations)

Products of Human Activities

Human Activities

The core conceptualization of Earth Sciences

Species (bio, geo)

Activities

Records

Observations

from cross reference exemplifies describe use to appear in Samples or Specimen exemplifies collections part of databases

Publications

Simulation

Place Time

is about global indices services

Forecasts

Complex System Situation occurs in create maintain from mathem.

Models

Products of Human Activities
The big picture (core concepts and relations)

Molecular world and parts

Species (bio,geo)

Activities

Observations

Simulation

Complex System

Samples or Specimen

collections

from

Place

Time

Situation

Math. Models

Records

Forecasts

databases

global indices

Publications

services

is about

use to appear in

cross reference

exemplifies

from

appear in

take, treat

use

about

happen in

observes

based on

describe

Yannis Tzitzikas, BlueBRIDGE Webinar, June 28, 2017
Knowledge Bases

The vision is to reach this area

Adequate for scientific needs

Simple or Naive

Conceptual modeling

specific domain

Scope

wide (or cross-domain)

The British Museum

GRSF

GeoNames

BIO22RDF

MusicBrainz

WordNet

BabelNet

DBpedia

yago

Eurostat - Linked Data
Concluding Remarks

• **Semantic integration** could boost data-intensive scientific discovery but requires tackling several challenging issues.
• We have discussed the main **difficulties, requirements and challenges** for semantic integration of data.
• We have presented a **process** and related **tools** that we have developed for supporting the **process of semantic integration in the marine domain** with emphasis on **Scope control, Ontological modeling, Connectivity assessment, Provenance, Reconstructability, Extensibility**
• Currently our research focuses on methods and techniques for semantically integrating **large number of datasets**.
• **MatWare** (for automating the warehouse construction process)
  • [http://www.ics.forth.gr/isl/MatWare/](http://www.ics.forth.gr/isl/MatWare/)

• **MarineTLO** (top-level ontology)

• **Semantic Warehouses**

• **XSearch** (exploiting semantic warehouses in searching)

• **Xlink** (exploiting semantic warehouses for entity identification in texts)
Links (2/2)

• Hippalus: Preference-enriched Faceted Search
  • [www.ics.forth.gr/isl/Hippalus](http://www.ics.forth.gr/isl/Hippalus)
    • Select a dataset from the Marine Biology domain for enacting the species identification through PFS

• GRSF VRE
  • [https://services.d4science.org/group/grsf](https://services.d4science.org/group/grsf)

Research Prototypes

• Interactive 3D Visualization of the LOD Cloud

• LODSyndesis (for measuring the commonalities in the entire LOD and offering global scale services)
  • [www.ics.forth.gr/isl/LODSyndesis](http://www.ics.forth.gr/isl/LODSyndesis)
Acknowledgements

• Joint work with
  • Yannis Marketakis
  • Nikos Minadakis
  • Michalis Mountantonakis
  • Pavlos Fafalios
  • Martin Doerr
  • Chryssoula Bekiari
  • Maria Papadaki
  • Panagiotis Papadakos
Thank you for your participation and attention

There is time for questions and discussion