Theophrastus: A Semantic Exploration Tool for Marine Taxonomists

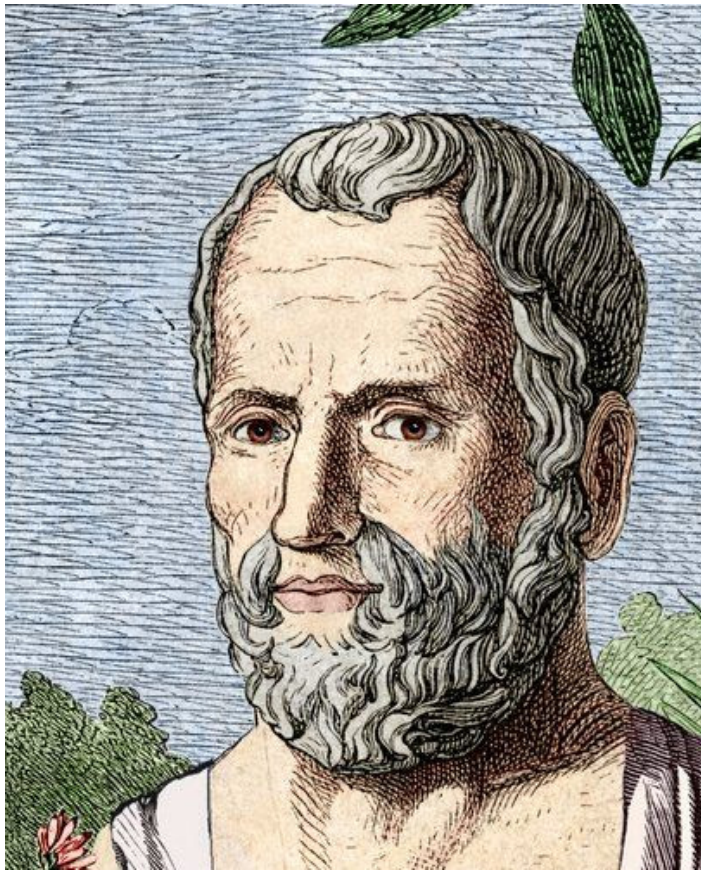
Fafalios Pavlos, Papadakos Panagiotis
{fafalios|papadako}@ics.forth.gr

FORTH-ICS & University of Crete

Blue Hackathon
July 2, 2013

Theophrastus

- Theophrastus (Greek: Θεόφραστος; c. 371 – c. 287 BC), was the successor to Aristotle in the Peripatetic school
- After Aristotle death, he continued his ichthyological research
- He composed a treatise on amphibious fish
- He offered the first systemization of the botanical world
- **Phoenix theophrasti**, the Cretan Date Palm



Data Available to Taxonomists

Many sources with available interesting data for taxonomists

- SPARQL endpoints for querying online Knowledge bases
- Web sources with literature related to a species
- Sources that provide available synonyms of a species in the literature

Taxonomists

Taxonomist Problematic Workflow

- **Diversity of Sources:**Taxonomists have to search for available information from a number of different places
- **Ambiguous Synonyms:** Available synonyms are sometimes ambiguous (due to the huge 250 years bibliography)
- **Disengagement:** Taxonomists have to focus on other information sources instead of the original one
- **Time Consuming:** Process is time consuming

Need

Taxonomists Need An Exploration Tool which:

- given a textual source (pdf, html) provides available information from **a number of different Knowledge Bases** in **real-time**
- when the taxonomist asks for it (i.e. **sparql query cost paid only when needed)**
- **annotates** the original information source
- is **generic** by configurable SPARQL endpoints and queries
  - take advantage of most accurate endpoints
  - support other areas of interest
- can **semantically enrich textual information** in RDFa
  - public repositories of RDFa enriched documents
  - search engines can take advantage of them (google, yahoo)
- taxonomist is **not disengaged** from the original information source

Data + Need = Hack (Theophrastus)

- **Entity mining** of original information source using GATE[1] (over PDF or html)
- Taxonomists provide **entities of interest (categories)** through a configuration file (i.e. species, water areas, countries, etc) (a SPARQL endpoint for each category)
- Associate each category with a number of **information needs** (i.e. species synonyms, species taxonomy, species belonging to the same family or genus, etc)
- Each **information need** is associated with a specific SPARQL query
- **Annotate** original document with recognized entities (could use different colors for different categories) (for PDF file the entities are displayed in a sidebar)
- **Enrich** original information source with **semantic information**, i.e. RDFa (not for PDF files)
- **Navigate** to related information gathered from queries (i.e. related species, etc) through pop-up windows
- All the above in **real-time**

[1] http://gate.ac.uk/ie/

Theophrastus DEMO

Thank you!