Quantifying the Connectivity of a Semantic Warehouse

Y. Tzitzikas[1,2], N. Minadakis[1], Y. Marketakis[1], P. Fafalios [1,2] ,
C. Alloca[1] and M. Mountantonakis[1,2]

[1] Institute of Computer Science, FORTH-ICS
[2] Computer Science Department, University of Crete, GREECE

FORTH Institute of Computer Science

4th International Workshop on Linked Web Data Management (LWDM),
Athens, May 28, 2014

Outline

- Motivation
- Context
- The process of constructing Semantic Warehouses
- The Connectivity Metrics
  - and results from using them in an operational semantic warehouse
- Concluding Remarks

Motivation

In many applications one has to fetch and assemble pieces of information coming from more than one sources (including SPARQL endpoints.)

Def: We use the term **Semantic Warehouse** (for short warehouse) to refer to a **read-only set of RDF triples fetched (and transformed) from different sources that aims at serving a particular set of query requirements**.

We focus on the aspects of quality and value (of the query answering) of the warehouse.

Key Questions

- *How to measure the value and quality (since this is important for e-science) of the warehouse?*
- *How to monitor its quality after each reconstruction or refreshing (as the underlying sources change)?*

The aspect of Connectivity

- In general, connectivity concerns both **schema** and **instances,** and it is achieved through common URIs, common literals and equivalence relations (e.g. `sameAs`)

- 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

Context

Context: iMarine

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

Id: It is 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 FIAT PANIS

IRD Institut de recherche pour le développement

HELLENIC REPUBLIC National and Kapodistrian University of Athens

CERN

FIN

UNESCO

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

ENGINEERING

terradue 2.0

ERCIM

Trust-IT Services Ltd Communicating ICT to markets

FORTH Institute of Computer Science

Marine 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: **in several sources**

Storing **complementary 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: in several sources

Using and accessed through **different technologies**

Web services (SOAP/WSDL)

RDF + OWL files

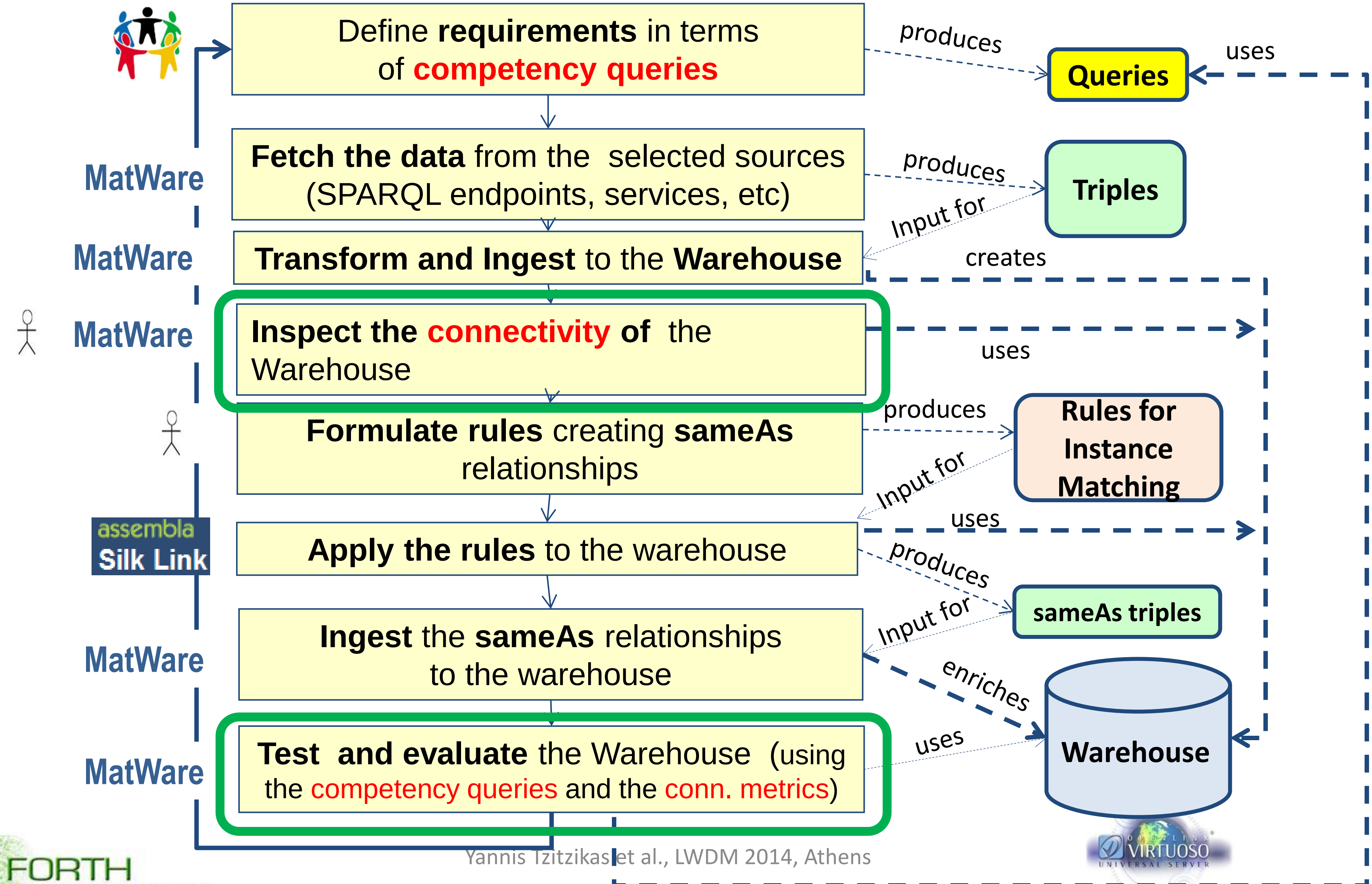
SPARQL Endpoint

Relational Database

SPARQL Endpoint

The MarineTLO-based semantic warehouse

The Warehouse construction and evolution process



FORTH Institute of Computer Science

The Metrics

Notations and Preliminaries

- $S_1 \ldots S_k$: the underlying sources
- $triples(S_i)$: the triples that S_i contributes to the warehouse W
- U_i: the URIs in the triples in $triples(S_i)$
- Lit_i: the literals in the triples in $triples(S_i)$

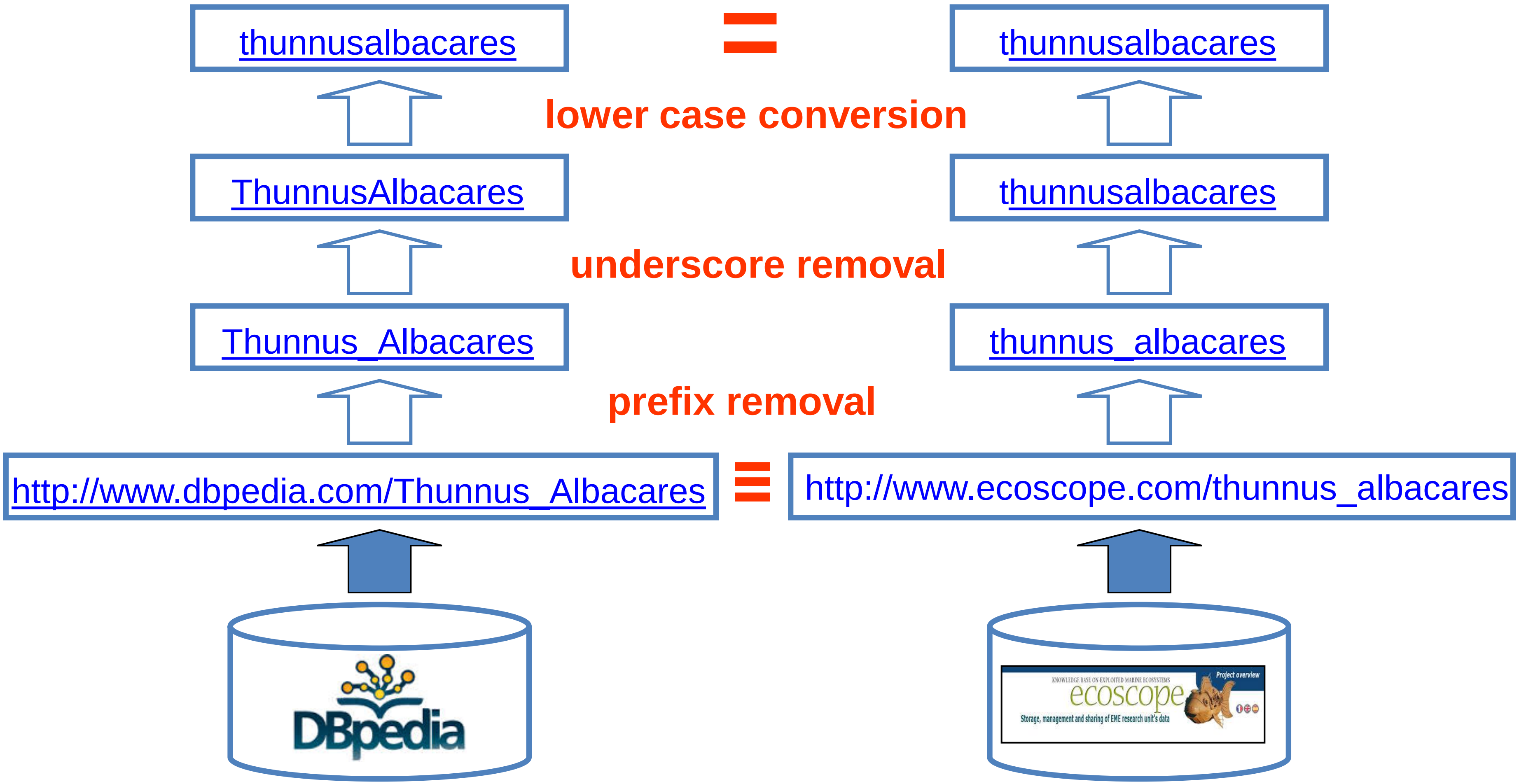
How to compare two sets of URIs, e.g. U_1 and U_2?

- There are more than one methods
- We propose the following three methods (policies)
  - The metrics that will be introduced can be computed using any of these policies

Policy Name	Policy Description
Exact String Equality	$u_1 = u_2 \Rightarrow u_1 \equiv u_2$
Suffix Canonicalization	$last(u_1) = last(u_2) \Rightarrow u_1 \equiv u_2$
Entity Matching	$u_1 \texttt{ sameAs } u_2 \Rightarrow u_1 \equiv u_2$

FORTH Institute of Computer Science

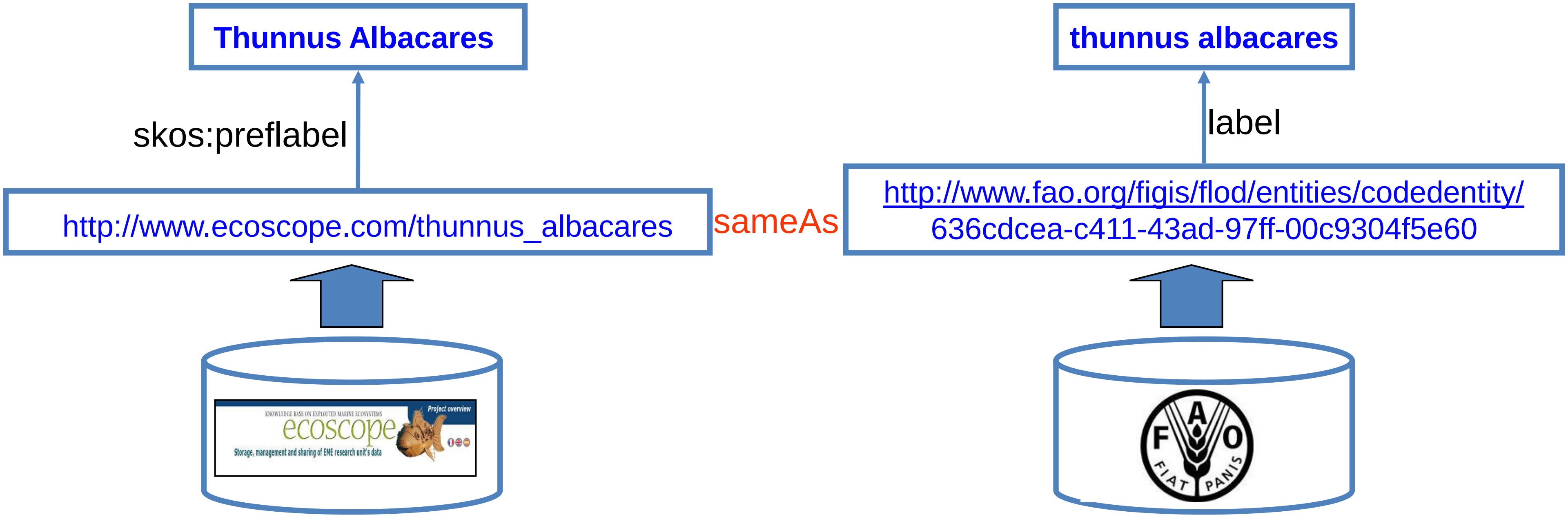
Example: **Suffix-based** URI equivalence



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.

Example: **Entity Matching**-based URI Equivalence

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.



Connectivity Metrics

- Proposed Metrics
  1. the **matrix of percentages of the common URIs**
  2. the **matrix of percentages of the common literals**
  3. the **increments in the average degree** of each source
  4. the **unique triple contribution** of each source
  5. the **complementarity factor** of the entities of interest

Metric 1 : Matrix of Percentages of Common URIs

$$curi_{i,j} = \frac{|U_i \cap U_j|}{\min(|U_i|,|U_j|)}$$

The percentage of common URIs between source S_i and S_j

Suffix canonicalization

S_i \ S_j	FLOD	WoRMS	Ecoscope	DBpedia	FishBase
FLOD	173,929	239	523	631	887
WoRMS		80,485	200	1,714	3,596
Ecoscope			5,824	192	225
DBpedia				70,246	9,578
FishBase					34,974

Common URIs ($|U_i \cap U_j|$)

S_i \ S_j	FLOD	WoRMS	Ecoscope	DBpedia	FishBase
FLOD	1	0.3%	8.98%	0.9%	2.54%
WoRMS		1	3.43%	2.44%	10.28%
Ecoscope			1	3.3%	3.86%
DBpedia				1	27.39%
FishBase					1

Common URIs % ($curi_{i,j} = \frac{|U_i \cap U_j|}{\min(|U_i|,|U_j|)}$)

Entity Matching

S_i \ S_j	FLOD	WoRMS	Ecoscope	DBpedia	FishBase
FLOD	190,733	434	1,897	4,009	6,732
WoRMS		80,486	805	1,754	3,596
Ecoscope			7,805	1,245	2,116
DBpedia				74,381	10,385
FishBase					34,974

Common URIs ($|U_i \cap U_j|$)

S_i \ S_j	FLOD	WoRMS	Ecoscope	DBpedia	FishBase
FLOD	1	0.54%	24.3%	5.39%	19.25%
WoRMS		1	10.31%	2.36%	10.28%
Ecoscope			1	15.95%	27.1%
DBpedia				1	29.69%
FishBase					1

Common URIs % ($curi_{i,j} = \frac{|U_i \cap U_j|}{\min(|U_i|,|U_j|)}$)

Metric 2 : Matrix of Percentages of Common Literals

$$clit_{i,j} = \frac{|Lit_i \cap Lit_j|}{\min(|Lit_i|, |Lit_j|)}$$

The percentage of common Literals between source S_i and S_j

S_i \ S_j	FLOD	WoRMS	Ecoscope	DBpedia	FishBase
FLOD	1	7.1%	12.37%	5.1%	8.55%
WoRMS		1	2.71%	4.76%	9.34%
Ecoscope			1	2.76%	2.99%
DBpedia				1	11.33%
FishBase					1

Common Literals % $(clit_{i,j} = \frac{|Lit_i \cap Lit_j|}{\min(|Lit_i|,|Lit_j|)})$

Metric 3 : Increase in the Average Degree

It shows the increment of the graph-theoretic degree of each entity when it becomes part of the warehouse graph.

$$\frac{\deg_W(E) - \deg_S(E)}{\deg_S(E)}$$

where

$$\deg_S(E) = avg_{e \in S}(|\{(s,p,o) \in S \mid s = e \;\; or \; o = e\}|)$$

FORTH Institute of Computer Science

Metric 3 : Increase in the Average Degree

S_i	avg $deg_{S_i}(U_i)$	avg $deg_W(U_i)$	increase
FLOD	7.18	9.18	27.84%
WoRMS	3.3	7.33	122.36%
Ecoscope	22.84	31.18	36.56%
DBpedia	41.41	42.11	1.7%
FishBase	18.86	29.81	58.08%
AVERAGE	18.72	23.92	27.78%

Average degrees in sources and in the warehouse

≡ Suffix canonicalization

The average degree is increased from 18.72 to 23.92.

S_i	avg $deg_{S_i}(U_i)$	avg $deg_W(U_i)$	increase
FLOD	7.18	54.31	656.51%
WoRMS	3.3	9.93	201.36%
Ecoscope	22.84	165.24	623.6%
DBpedia	41.41	84.2	103.36%
FishBase	18.86	50.6	168.32%
AVERAGE	18.72	72.86	289.21%

Average degrees in sources and in the warehouse

≡ Entity Matching

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

Metric 4 : Unique Triple Contribution

$$triplesUnique(S_i) = \frac{triples(S_i)}{\bigcup_{1 \le j \le k, i \ne j} triples(S_j)}$$

It shows the unique triple contribution of each source, which are the number of triples for each source excluding triples that provided by any other source.

FORTH
Institute of Computer Science

Metric 4 : Unique Triple Contribution

- Suffix canonicalization

S_i	$a = \|triples(S_i)\|$	$b = \|triplesUnique(S_i)\|$	b/a
FLOD	665,456	664,703	99.89%
WoRMS	461,230	460,741	99.89%
Ecoscope	54,027	53,641	99.29%
DBpedia	450,429	449,851	99.87%
FishBase	1,425,283	1,424,713	99.96%

(Unique) triple contributions of the sources

- Entity Matching (and ingestion transformations)

S_i	$a = \|triples(S_i)\|$	$b = \|triplesUnique(S_i)\|$	b/a
FLOD	810,301	798,048	98.49%
WoRMS	582,009	527,358	99.88%
Ecoscope	138,324	52,936	38.27%
DBpedia	526,016	517,242	98.33%
FishBase	1,425,283	1,340,968	94.08%

(Unique) triple contributions of the sources

Metric 5 : Complementarity Factor

$$cf(e) = |\{i \mid triples_W(e) \cap triplesUnique(S_i) \neq \emptyset\}|$$

The complementarity factor of the entities of interest is the number of sources that provided unique triples for each entity of interest

*(with the term **entity** we mean any literal or URI that contains the corresponding entity name, e.g the string "thunnus")*

For the entities *Thunnus* and *Shark*, all the sources provided unique triples, but for the entities *Greece* and *Astrapogon* only three sources provided unique material.

Kind of Entity	$cf(\cdot)/5$
Thunnus	5/5
Greece	3/5
Shark	5/5
Astrapogon	3/5

Complementarity factor (cf) of some entities

Detecting Redundancies or other Pathological Cases

- The metrics allow spotting pathological cases e.g. redundant sources or totally unconnected sources
- We defined two artificial sources
  - **CloneSource:** a subset of Ecoscope's and DBpedia's triples as they are stored in the warehouse.
  - **Airports:** containing triples about airports which were fetched from the DBpedia public SPARQL endpoint
- Results
  - **CloneSource**: 0 unique contribution as expected, since it was composed from triples of existing sources
  - **Airports**: The increase in the average degree for the entities of that source was very low (due to some common country names)
- General Rules for identifying problematic cases
  - **1)** If the unique contribution of a source is very low (resp. zero), then this means that it does not contribute significantly (resp. at all) to the warehouse.
  - **2)** If the average increase of the degree of the entities of a source is low, then this means that its contents are not connected with the contents of the rest sources.

Metrics Results Displayed In HTML as computed by MatWare

Metrics Results

Produced by MaTWare on: 1/12/2013
SPARQL EndPoint: http://virtuoso.i-marine.d4science.org:8890/sparql
Sources Used: i)FLOD ii)WoRMS iii)Ecoscope iv)DBpedia v)Fishbase vi)Clone Source vii)Airports

Common Uris

Source	FLOD	WoRMS	Ecoscope	DBpedia	Fishbase	Clone Source	Airports
FLOD	173929	239	523	631	887	250	13
WoRMS		80485	200	1714	3596	364	0
Ecoscope			5824	192	225	4030	4
DBpedia				70246	9578	4589	14
Fishbase					34974	481	60
Clone Source						8457	4
Airports							4606

Common Uris Percentage

Source	FLOD	WoRMS	Ecoscope	DBpedia	Fishbase	Clone Source	Airports
FLOD	1	0.3%	8.98%	0.9%	2.54%	2.96%	0.28%
WoRMS		1	3.43%	2.44%	10.28%	4.3%	0%
Ecoscope			1	3.3%	3.86%	69.2%	0.09%
DBpedia				1	27.39%	54.26%	0.3%
Fishbase					1	5.69%	1.3%
Clone Source						1	0.09%
Airports							1

Common Literals

Source	FLOD	WoRMS	Ecoscope	DBpedia	Fishbase	Clone Source	Airports
FLOD	111164	3624	1745	5668	9504	373	1533
WoRMS		51076	382	2429	4773	289	86
Ecoscope			14102	389	422	6871	131
DBpedia				123887	14038	7144	117
Fishbase					138275	604	152
Clone Source						13964	49
Airports							12302

Common Literals Percentage

Source	FLOD	WoRMS	Ecoscope	DBpedia	Fishbase	Clone Source	Airports
FLOD	1	7.1%	12.37%	5.1%	8.55%	2.67%	12.46%
WoRMS		1	2.71%	4.76%	9.34%	2.07%	0.7%
Ecoscope			1	2.76%	2.99%	49.21%	1.06%
DBpedia				1	11.33%	51.16%	0.95%
Fishbase					1	4.33%	1.24%
Clone Source						1	0.4%
Airports							1

Triples

Source	Triples	Unique Triples	Percentage
FLOD	665456	664703	99.89%
WoRMS	461230	460741	99.89%
Ecoscope	54027	17951	33.23%
DBpedia	450429	429426	95.34%
Fishbase	1425283	1424713	99.96%
Clone Source	56166	0	0%
Airports	31628	31628	100%

* Probably redundant source

Complementarity Factor

Entities	Complementarity Factor
Astrapogon	2/7
Species	5/7
Greece	4/7
Thunnus	5/7
Shark	5/7

Degrees

Source	Source Degree	Warehouse Degree	Increase
FLOD	7.18	54.3	656.4%
WoRMS	3.3	9.93	200.09%
Ecoscope	22.84	165.24	623.46%
DBpedia	41.41	84.2	104.8%
Fishbase	18.86	50.6	168.29%
Clone Source	44.43	84.2	89.5%
Airports	70.99	72.56	2.2%
Average	41.8	74.43	78.07%

* Probably out of domain of interest

FORTH Institute of Compu

Concluding Remarks

- We have proposed metrics for quantifying the connectivity of a semantic warehouse:
  1. the **matrix of percentages of the common URIs and/or literals**,
  2. the **complementarity factor** of the entities of interest,
  3. the **increments in the average degree** of each source, and
  4. the **unique triple contribution** of each source.
- By inspecting the proposed metrics-based matrixes one can very quickly get an **overview** of the contribution of each source and the tangible **benefits** of the warehouse
- The values of (1),(2),(3) allow evaluating the warehouse, while (3) and (4) mainly concern each particular source.
- One can exploit these metrics:
  - Before adding a new source: for investigating if a source is interesting for importing in the warehouse (e.g. to inspect if this source contributes unique information)
  - After a warehouse update: for inspecting and controlling the evolution of the warehouse, for inspecting if an update of a source affected positively or negatively the connectivity of the warehouse
- We are currently working on extension of VoID for expressing and exchanging in machine processable (and query-able) manner these metrics

Thank you for your attention

Visit and send us feedback:

www.ics.forth.gr/isl/MarineTLO

www.ics.forth.gr/isl/MatWare