# MatWare: A Framework for Automating the Creation of Semantic Warehouses

Overview

In many applications one has to fetch and assemble pieces of information coming from more than one web sources, such as SPARQL endpoints. For this reason ISL has developed MatWare (derived from Materialized Warehouse), a framework for constructing and evaluating domain-specific semantic warehouses. **Matware automates**: (a) the process of constructing a warehouse (downloading of sources, creation of sameAs links, etc.), (b) the refreshing of the warehouse as data derived from the underlying sources change and (c) the quantitative evaluation of the resulted warehouse by supporting a novel set of metrics.

MatWare is framework that automates the process of constructing semantic warehouses. We use the term semantic 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. MatWare automatically fetches contents from the underlying sources using several access methods (SPARQL endpoints, HTTP accessible files, JDBC). The fetched data are stored (in their original form or after transforming them) in a RDF triplestore. To tackle the **provenance requirements** several levels are supported: (a) URIs and values level, (b) triple level (separate graphspaces) and (c) query level. MatWare also supports the refreshing of the warehouse by replacing specific sources as they evolve (instead of re-constructing the whole warehouse from scratch).

The process of constructing and evolving the warehouse

*An Overview of the MarineTLO-based warehouse*
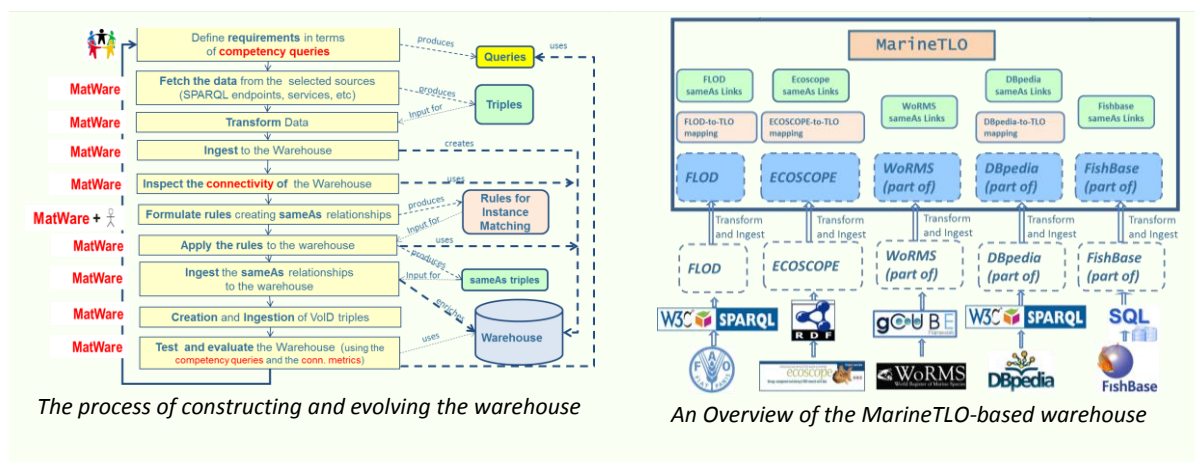
Target Domains

MatWare can be used for constructing both general purpose semantic warehouses and domain-specific warehouses. In the context of iMarine project (FP7 Research Infrastructures), MatWare is employed for constructing warehouses for the biodiversity domain.

Description

A distinctive feature of MatWare is that it allows **evaluating the connectivity of the warehouse**. We use the term connectivity to express the degree up to which the contents of the semantic warehouse form a connected graph that can serve, ideally in a correct and complete way, the query requirements of the semantic warehouse, while making evident how each source contributes to that degree. To this end, MatWare supports several metrics. These metrics allow someone to get an overview of the contribution (to the warehouse) of each source (enabling the discrimination of the important from the less-important sources) and to quantify the benefit of such a warehouse. These metrics include matrixes of common URIs/literals between different sources, percentages of the unique contribution of triples of each source, average degree of the entities of each source, complementarity factors (e.g. the number of sources that provided unique triples for the entities of interest), etc.
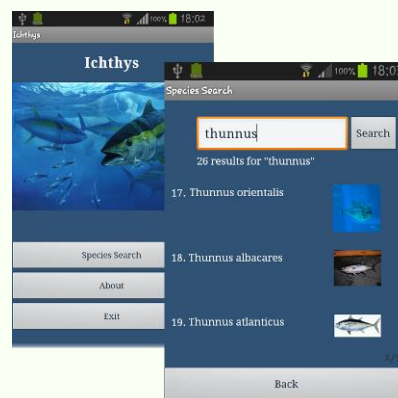
Screenshots of the "Ichthys" Android application
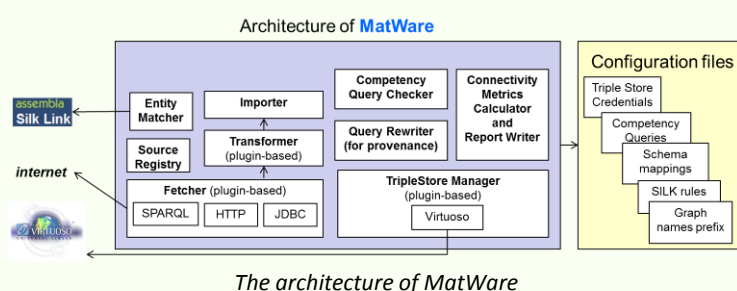
The architecture of MatWare

Source	Triples	Unique Triples	Percentage
Ecoscope	136617	51305	37.55%
Worms	692535	691414	99.84%
Fishbase	1270658	1186341	93.36%
DBpedia	529566	520496	98.29%
FLOD	799305	788347	98.63%

A screenshot from the evaluation of the warehouse indicating the unique contribution of each source in the warehouse

Additional Information

MatWare was partially **supported** by the FP7 EU project iMarine and H2020 EU project BlueBRIDGE. The resulting warehouses are **used** by several applications including: meta-search engines (i.e. XSearch) that support the semantic post-processing of search results, mash-up applications, and mobile applications (i.e. Ichthys).

More **information** about MatWare can be found at *http://www.ics.forth.gr/isl/MatWare*

MatWare website

Contact details: **Yannis Tzitzikas**

tzitzik@ics.forth.gr

www.ics.forth.gr/isl