A Reconciliation Framework for the Integration of Stocks and Fisheries Information

Yannis **Marketakis** [1], Yannis **Tzitzikas** [1,2], Aureliano **Gentile** [3], Bracken **van Niekerk** [3] and Marc **Taconet** [3]

[1] *Institute of Computer Science, FORTH-ICS, Heraklion, Greece*
[2] *Computer Science Department, University of Crete, Heraklion, Greece*
[3] *Food and Agriculture Organization of the United Nations, Rome, Italy*

Abstract
Fisheries management relies on analyzing data using complex models and software and includes the usually manual process of identifying and combining different parts of information about stocks and fisheries, which is a time-consuming and error-prone process. Firstly because there is no single source of information but rather they are many, and secondly because there are alternative ways of modeling and referring to the same piece of information. Approaches like the Global Record of Stocks and Fisheries (GRSF), which are the result of the semantic data integration of the corresponding information from different data sources, aim to overcome such problems, by providing a unified view of the stocks and fisheries information in a homogeneous manner. In this paper, we propose a reconciliation framework ensuring that similar pieces of information from heterogeneous sources are properly connected during the construction of the semantic warehouse of GRSF.

1. Introduction

The main goal of fisheries management is to monitor, specify and propose regulations and rules for protecting the fishery resources, so that their sustainable exploitation is possible. Although there is no clear and generally accepted definition of fisheries management, according to [1] the main task of fisheries management is the integrated process of information gathering, analysis, planning, consultation, decision-making, allocation of resources and formulation and implementation, with enforcement as necessary, of regulations or rules which govern fisheries activities in order to ensure the continued productivity of the resources and the accomplishment of other fisheries objectives. The key indicators for efficient fisheries management are stocks and fisheries. Stocks refer to groups or individuals of a species occupying a well-defined spatial range (e.g. *swordfish* in the Mediterranean Sea), while fisheries describe the activities leading to the harvesting of the fish within a particular area, using a particular method or equipment and purpose of activity (e.g. the *Atlantic cod* fishery in the area of East and South Greenland).

Nowadays there are several stakeholders, reporting information about stocks and fisheries at regional, national and local levels. These stakeholders maintain their knowledge and publicize their contents independently from each other. However, cross-references between the knowledge bases from different stakeholders is not a common issue. In addition, the use of common vocabularies or standards

for describing particular aspects has not been globally agreed. This leads to the creation of several data silos, each one with its own rules and terminology.

The Global Record of Stocks and Fisheries (GRSF) [5], aims to overcome such problems by introducing a workflow that collects and semantically integrates stocks and fisheries information from different databases, and present them in a unified manner. More specifically, it is the result of the integration of (a) FIRMS [2], (b) RAM [3] and (c) FishSource [4]. During the construction and refreshment [6] of GRSF, a reconciliation workflow is applied, ensuring that if there are similar resources expressed in a different way they will be properly linked. In the following of the paper we describe such cases. More specifically, in Section 2, we further discuss about the problem, in Section 3 we elaborate with the reconciliation framework and its applicability, and Section 4 concludes our work.

2. Motivation

Since the sources contributing stocks and fisheries information to GRSF contain complementary information, GRSF provides a merged view of those records. Merging is applied by comparing some key elements of the records. For the case of stock records, the elements that are compared are: species and assessment area, while for the case of fisheries they are: species, fishing area, management authority, fishing gear, flag state. It becomes evident that the accuracy of the information of these elements, specify the successful merging or not.

Of course, there is not a single way of describing those elements. For example, marine species can be referred to using their scientific name (e.g. *Thunnus albacares*), their common name in any language (e.g. *Yellowfin tuna* in English), their 3-Alpha code (YFT), their APHIA ID (e.g. 127027)), etc. In a similar manner, an area can be referred to using a common name, a FAO major area code (e.g., 37.3.1), a GFCM code, an LME code, an ISO3 code of the exclusive economic zone, etc. The same applies for all the aforementioned elements.

Moreover, it is quite common that different sources use different terminologies for describing their contents. For example, in FIRMS species are usually referred to using their 3-Alpha codes, RAM reports their scientific name and taxonomy, and FishSource uses various combinations. It is therefore evident that before applying any merging activity in GRSF, it is required to build a knowledge base, with the different identifiers of each resource, so that they can be efficiently compared.

3. Reconciliation Framework

The reconciliation framework is responsible for constructing a knowledge base consisting of the different identifiers of the entities. The knowledge base can be afterwards exploited during the construction of the GRSF, so that (a) the merging of records from different sources can be executed properly and (b) all the records in GRSF use the preferred terminologies with respect to GRSF guidelines. Below we introduce the conceptual backbone of the constructed knowledge base, we describe the reconciliation framework workflow and we provide more information about the applicability of the framework for the purposes of GRSF.

3.1. Model

The configuration of the conceptual model for the reconciliation framework of GRSF is shown in the upper part of Figure 1. The core element is *Term*, and is associated with the elements *Identifier* and *Information Object*. More specifically, a Term can have a preferred identifier, one or more alternative identifiers, and many additional information objects. In order to compare two different terms, the identifier elements are used. In particular, if two terms have at least one of their identifiers the same then they are merged into one. For this reason, we are comparing all the available identifiers and a match is confirmed if all the values of an identifier match (i.e. *identifier*, *type*, *index*). For terms not having any identifier, we are comparing their additional information objects. The main difference is that when two terms have the same additional information object (same *value* and *type*), and at least one of them does not have any identifier, a suggestion of merging the terms is made and is it up to the

user to approve it. The lower part of Figure 1 shows an indicative example for the term about the species with scientific name "*Thunnus albacares*".
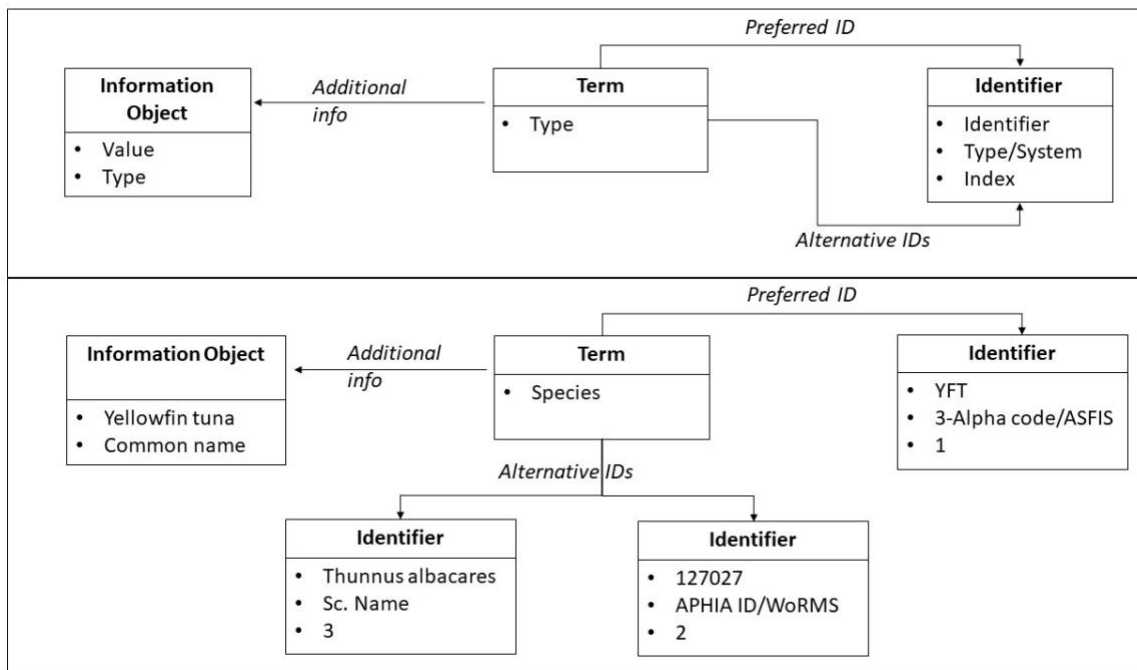

Figure 1: The Reconciliation framework data model

Identifier elements are enough for comparing and merging two terms. Practically, this means that as soon as two terms have at least one of their identifiers the same then they are matched, no matter if the identifier is the preferred or an alternative one. Furthermore, each identifier is associated with an index, representing the order of preference. Upon the construction of all the terms, they can be used for selecting the most preferred identifier of a term, starting from the preferred identifier (that has always index =1), and if that does not exist, the alternative identifier with the lowest index is used. In the example of Figure 1 the preferred identifier is YFT (3-Alpha code), the second most preferred is 127027 (APHIA ID), and so on. The types and indexes of the identifiers are specified during the configuration of the GRSF construction/refresh workflow. An indicative configuration of the types and indexes of the identifiers for GRSF is given in Section 3.3.

3.2. Workflow

Figure 2 shows the overall workflow that is used for constructing GRSF and the placement of the reconciliation workflow. The detailed description of the construction workflow and the refreshment of GRSF can be found in [5] and [6]. More specifically, the reconciliation KB is constructed using as input the transformed data as well as data from external sources. For each concrete resource, a new entry is added in the knowledge base with the available information. Of course it is expected that some information might be missing during the creation of a new terminology resource. The reconciliation framework will take care of complementing the missing information as soon as they exist somewhere else (e.g. in another record, a transformed source, or an external source). Particularly, if a term with a particular identifier already exist, then its information are complemented with the new ones, otherwise a new term is created.
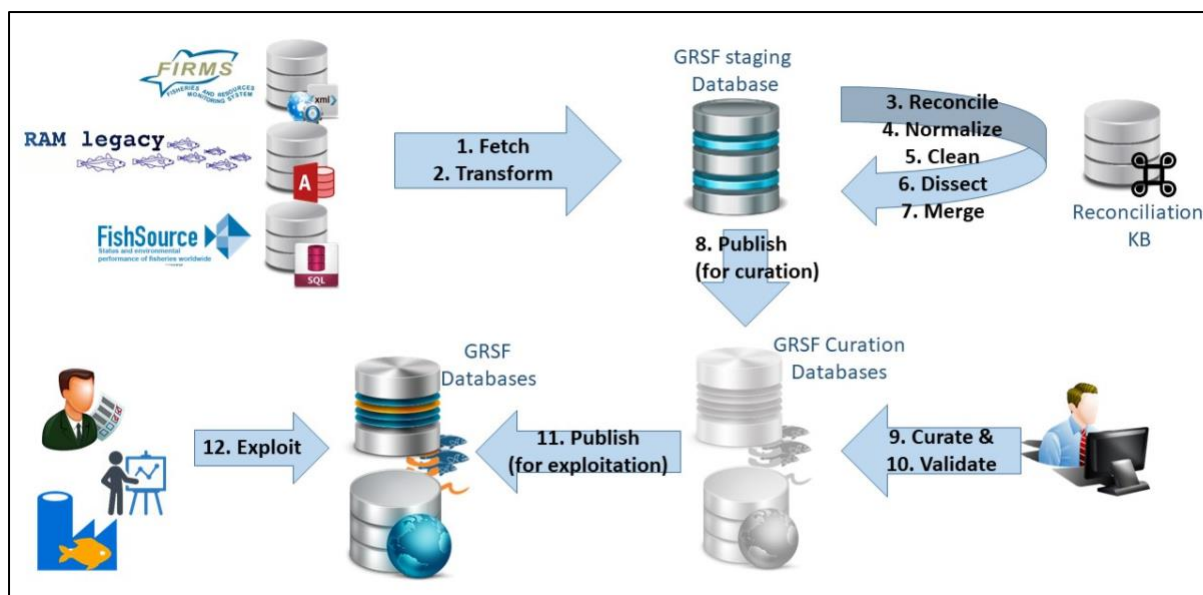
Figure 2: The GRSF construction workflow

Figure 3 illustrates an example depicting three stock records, occupying the same species, and each source documents that species differently. S1 will create a new term of entity type Species, with preferred identifier "*YFT*", and additional identifier "*Thunnus albacares*" with type scientific name. S2 instead of creating a new term, it will enhance the already created term shown before, because they have the same additional identifier of type scientific name. Finally, S3 will further enhance the term with an additional identifier.
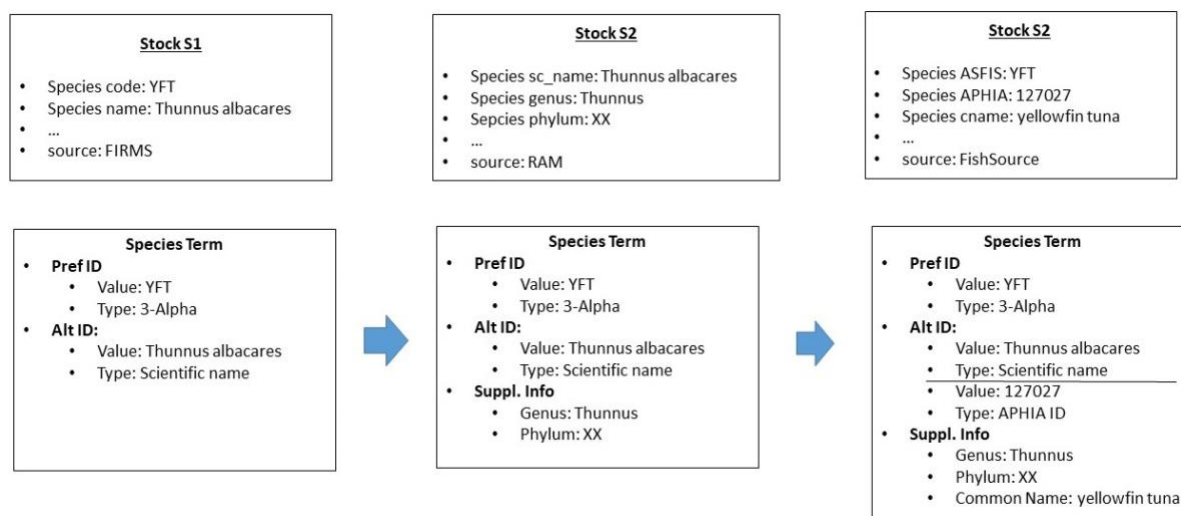

Figure 3: Creating and enhancing a term

After the construction of the reconciliation KB, it is exploited during the construction of the GRSF records. This is mandatory, so that records are fully compliant with the GRSF guidelines, that specify how the different information of GRSF records should be presented. This is achieved through the lookup methods, that allow spotting terms from the KB. As soon as a term is located, its preferred identifier is used for the construction of the corresponding GRSF record. If it does not exist, the alternative identifier with the lowest index will be used.

3.3. Application

Table 1 shows the list of term types that are reconciled using the reconciliation framework for the purposes of GRSF. For each term type, we provide the preferred identifier, and the alternative ones, as they have been agreed in GRSF guidelines.

Table 1 The term types and their identifiers

Term Type	Preferred ID / Index	Alternative IDs / Index
Species	3-Alpha[7] code / 1	APHIA[8] ID / 2
		Scientific Name / 3
Areas (assessment, fishing)	FAO[9] code / 1	GFCM[12] / 2
		LME[10] / 3
		MRGID[11] / 4
		ISO-3[13] (EEZ) / 5
Fishing Gear	ISSCFG[14] code / 1	ISSCFG abbrev. / 2
		ISSCFG category / 3
Flag State	ISO-3 / 1	Legal name / 2
Management Authority	Acronym / 1	Full name / 2

4. Conclusion

This paper demonstrates the reconciliation framework that facilitates the semantic data integration of data sources with stocks and fisheries and is compliant with the GRSF guidelines and best practices. Although we focus on the particular use case, the framework is generic enough and can be used across different domains as well. In fact, in this paper we have described its configuration on top of the GRSF construction workflow, with the objective of efficiently comparing and locating similar entities that have been described using different identifiers.

5. Acknowledgements

6. References

[1] K. L. Cochrane and S.M. Garcia, A Fishery Manager's Guidebook, John Wiley & Sons, 2009.
[2] Fisheries and Resources Monitoring System, 2022. URL: http://firms.fao.org/firms/en
[3] The RAM Legacy Stock Assessment Database, 2022. URL: https://www.ramlegacy.org/
[4] Sustainable Fisheries Partnership, FishSource 2022. URL: https://www.fishsource.org/
[5] Y. Tzitzikas, Y. Marketakis, N. Minadakis, M. Mountantonakis, L. Candela, F. Mangiacrappa, P. Pagano, C. Perciante, D. Castelli, M. Taconet, A. Gentille, G. Gorelli, Towards a Global Record of Stocks and Fisheries, 8th International Conference on Information and Communication Technologies in Agriculture, Food & Environment, September 21-24, 2017, Chania, Crete, Greece.
[6] Y. Marketakis, Y. Tzitzikas, A. Gentille, Bracken Van Niekerk, and Marc Taconet, On the Evolution of Semantic Warehouses: The Case of Global Record of Stocks and Fisheries, 14th International Conference on Metadata and Semantics Research, Special Track on Metadata & Semantics for Agriculture, Food & Environment (MTSR'20) Madrid, 2020.
[7] ASFIS List of Species for Fishery Statistics Purposes. URL: https://www.fao.org/fishery/en/collection/asfis/en

[8] W. Appeltans, M.J. Costello, B. Vanhoorne, W. Decock, L. Vandepitte, F. Hernández, J. Mees, E. Vanden Berghe, E., 2008. Aphia for a World Register of Marine Species (WoRMS). VLIZ Special Publication.

[9] FAO Major Fishing Areas. URL: https://www.fao.org/fishery/en/area/search

[10] Large Marine Ecosystems (LME). URL: https://lmehub.net/

[11] Marine Regions. URL: https://www.marineregions.org/

[12] General Fisheries Commision for the Mediterranean (GFCM). URL: https://www.fao.org/gfcm/about/en/

[13] ISO 3166. URL: https://www.iso.org/iso-3166-country-codes.html

[14] The International Standard Statistical Classification of Fishing Gear (ISSCFG). URL: https://data.apps.fao.org/catalog/dataset/the-international-standard-statistical-classification-of-fishing-gear-isscfg