Methods and Tools for Supporting the Integration of Stocks and Fisheries

Yannis Tzitzikas ✉[1,2][0000-0001-8847-2130], Yannis Marketakis[1][0000-0002-0417-2526], Nikos Minadakis[1,2][0000-0001-6062-9298], Michalis Mountantonakis[1,2][0000-0002-1951-0241], Leonardo Candela[3][0000-0002-7279-2727], Francesco Mangiacrapa[3], Pasquale Pagano[3][0000-0001-6611-3209], Costantino Perciante[3][0000-0001-6056-8448], Donatella Castelli[3], Marc Taconet[4], Aureliano Gentile[4], Giulia Gorelli[4]

[1] Institute of Computer Science, FORTH-ICS Heraklion, Greece
{tzitzik,marketak,minadakn,mountant}@ics.forth.gr
[2] Computer Science Department, University of Crete, Heraklion, Greece
[3]Consiglio Nazionale delle Ricerche, Pisa, Italy
{leonardo.candela,francesco.mangiacrapa,pasquale.pagano,
costantino.perciante,donatella.castelli}@isti.cnr.it
[4]Food and Agriculture Organization of the United Nations, Rome Italy
{ marc.taconet,aureliano.gentile,giulia.gorelli}@fao.org

Abstract. The collation of information for the monitoring of fish stocks and fisheries is a difficult and time-consuming task, as the information is scattered across different databases and is modelled using different formats and semantics. Our purpose is to offer a unified view of the existing stocks and fisheries information harvested from three different database sources (FIRMS, RAM and FishSource), by relying on innovative data integration and manipulation facilities. In this paper, we describe the building blocks in terms of methods and software components that are necessary for integrating stocks and fisheries data from heterogeneous data sources.

Keywords: fish stock, fishery, semantic data integration, data publication, data normalization

1 Introduction

Fish Stocks are groups of individuals of a species occupying a well-defined spatial range independent of other stocks of the same species, e.g. swordfish in the Mediterranean Sea[1]. A ***Fishery*** is a unit determined by an authority or other entity that is engaged in raising and/or harvesting fish. Typically, the unit is defined in terms of some or all of the following: people involved, species or type of fish, area of water or seabed, method of fishing, class of boats and purpose of activity, e.g. Fishery for At-

[1] http://firms.fao.org/firms/resource/10025/en

lantic cod in the area of East and South Greenland[2]. Information about Fish Stocks and Fisheries is widely used for the monitoring of their status, and to identify appropriate management actions [1], with the ultimate goal of sustainable exploitation of marine resources. For these reasons completeness, adequacy and validity of information is crucial. Although this key role, there is no “one stop shop” for accessing stocks and fisheries data. Such information is usually collected (and produced as a result of data analysis) by the fishery management authorities at regional, national and local level. Therefore, the overall information is scattered across several databases, with no standard structure due to the specific local needs of the different bodies. Furthermore, the guidelines for populating existing registries are therefore heterogeneous, and every registry is actually a “database silo” that is not expected to interoperate with others to offer a global view on existing information.

Our objective (in the context of the ongoing BlueBRIDGE EU project[3]) is to construct a Global Record of Stocks and Fisheries (for short GRSF) capable of containing the corresponding information categorized into uniquely and globally identifiable records. Instead of creating yet another registry, we focus on producing GRSF records by using existing data. This approach does not invalidate the process being followed so far, in the sense that the organizations that maintain the original data are expected to continue to play their key role in collecting and exposing them. In fact, GRSF does not generate new data, rather it collates information coming from the different database sources, facilitating the discovery of inventoried stocks and fisheries arranged into distinct domains.

The advantages of this approach include: (a) increased data coverage compared to the single sources of information, (b) integrating information and identifying unique stocks and fisheries coming from the different database sources, and (c) answering queries that would be impossible to be answered from the individual database sources. These characteristics meet the needs of the main business cases that are: (i) supporting the compilation of stock status summaries at regional and global level and (ii) providing services for the traceability of sea-food products.

In this paper, we extend our previous work, described in [2]. In that work we described the methodology and the software components that were used for constructing GRSF, and presented some first results of the registry. In the current paper we focus on the methodology and the processes that were carried out for integrating heterogeneous information from the remote data sources. In addition, we describe the activities that were performed for normalizing the harvested and transformed data.

The rest of this paper is organized as follows. Section 2 describes the background information. More specifically it describes the main requirements and the data sources that were used for constructing GRSF. Section 3 discusses the data normalization activities, while Section 4 describes the technical framework that was used. Finally, Section 5 concludes and identifies directions for future work and research.

[2] https://www.fishsource.org/stock_page/688

[3] BlueBRIDGE Project website http://www.bluebridge-vres.eu/

2 Background

In this section, we summarize the basic information of GRSF, as they have been described in detail in our previous work [2]. More specifically we discuss about the data sources that were exploited (§ 2.1), the main requirements (§ 2.2), the structure of the final GRSF record (§ 2.3) and the overall process (§ 2.4).

2.1 The Data Sources

Below we describe the three database sources that have been used so far to harvest stocks and fisheries information. These sources are (a) Fisheries and Resources Monitoring System (FIRMS), (b) RAM Legacy Stock Assessment database, and (c) FishSource. The rationale for the selection of these sources, is that they contain complementary information (both conceptually and geographically). More specifically FIRMS is mostly reporting at regional level, while RAM is reporting at national or subnational level, and FishSource is more focused on the fishing activities. All of them contribute to overall aim to build a comprehensive and transparent global reference set of stocks and fisheries records that will boost regional and global stocks and fisheries status and trend monitoring as well as responsible consumer practices. Since the construction of GRSF is an iterative process, we will support integrating contents from these three sources in early releases of GRSF, and in future we will investigate exploiting new ones (i.e. FAO Global Capture Production Statistics database[4]).

FIRMS (FIsheries and Resources Monitoring System)[5] provides access to a wide range of high-quality information on the global monitoring and management of stocks and fisheries. It collects data from 14 intergovernmental organizations (that are partners of FIRMS) and contains information on the status of more than 600 stocks and 300 fisheries. The information provided by the organizations is organized in a database and published in the form of XML backboned fact sheets.

RAM (RAM Legacy Stock Assessment Database)[6] provides information exclusively on the fish stocks domain. It is a compilation of stock assessment results and time series of stock status indicators for commercially exploited marine populations from around the world. The assessments were assembled from 21 national and international management agencies for approximately one thousand stocks. RAM contents are stored in a relational database and are publicly available by releasing versions of the database in MS Access and Excel format.

FishSource[7] compiles and summarizes publicly available scientific and technical information about the status of fish stocks and fisheries. It includes information about the health of stocks, the quality of their management, and the impact of fisheries on the rest of the ecosystem. It is mainly exploited from seafood industry for assisting in

[4] http://www.fao.org/fishery/statistics/global-capture-production/en

[5] http://firms.fao.org/firms/en

[6] http://ramlegacy.org

[7] http://www.fishsource.com/

taking the appropriate actions for improving the sustainability of the purchased seafood. Information in FishSource is organized into fishery profiles associated with the exploited stocks. The database contains information for more than 2,000 fishery profiles.

2.2 Requirements

The selected database sources were constructed to fulfil different requirements and needs. Furthermore, they have been developed and are maintained from different initiatives. As a result, they are using different standards, data models, conceptualizations and terminologies for capturing similar information. As an example consider the fish species that are included in a particular stock or fishery; they can be identified either using (a) their scientific name (e.g. Thunnus albacares), (b) their common name in any language (e.g. Yellowfin tuna in English), or (c) standard codes for identifying them (e.g. YFT[8]). Furthermore, the different data sources use diverse criteria for identifying the uniqueness of a stock or fishery, as well as diverse conventions for naming their records.

GRSF aims at harmonizing the harvested information by adopting a set of standards that have been discussed and agreed with representatives of the database sources. In particular, these standards have been identified by two technical working group meetings that have been organized with the support of the BlueBRIDGE project. The working groups have defined which are the international standards that will be used (e.g. FAO 3Alpha codes for species, ISO3 country codes for flag states), which values define the uniqueness of a stock or a fishery record, which values are mandatory to accept a record as a complete one, as well as guidelines for generating unique and global identifiers (both human and machine interpretable) and names for the GRSF records. A detailed description of a GRSF record with respect to those guidelines can be found in Section 2.3 .

The main challenge for the construction of the GRSF is the ability to semantically integrate data coming from different data sources. To tackle this challenge, we decided to rely on semantic web technologies and use top level ontologies. The best candidate is the MarineTLO [3] which provides (a) consistent abstractions or specifications of concepts included in all data models or ontologies of marine data sources and (b) the necessary properties to make GRSF a coherent source of facts relating observational data with the respective spatiotemporal context and categorical domain knowledge. The rationale is that we map attributes from different data sources into classes and properties of the top level ontologies. To this end we could also mention works like [4] that automate the mapping process using machine learning techniques.

[8] According to FAO 3Alpha code http://www.fao.org/fishery/collection/asfis/en

2.3 The GRSF Record

Each GRSF record is composed of several fields to accommodate the incoming information and data. The fields can be functionally divided into time-independent and time-dependent. The first group contains the identification information which uniquely defines a stock or fishery, the latter contains the stocks and fishery indicators. In general, there are two types of GRSF records: (a) stocks and (b) fishery GRSF records. Both types of records share some common metadata like their identification details, and descriptive information. Furthermore, records are assigned information about areas and their original sources. Finally, each record is assigned several time-dependent information modeled as dimensions. In the case of stock GRSF records, the dimensions refer to abundance levels and exploitation rates. In the cases of fishery GRSF records, the dimensions refer to catches and landings indicators. We could say that a GRSF record resembles a data item in a database and as such we are describing its corresponding details in the schema shown in Fig. 1.

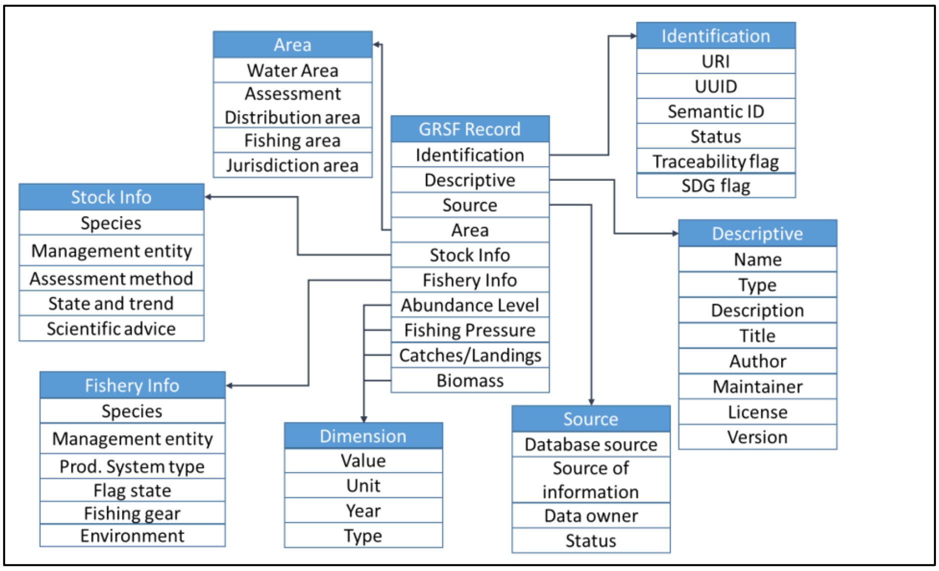

Fig. 1. The STAR schema of a GRSF record

2.4 The Process

The process for constructing GRSF consists of a sequence of steps which are shown in Fig. 2. Below we describe these steps in detail. More information about particular parts of the process (i.e. the data normalization and cleaning steps) are described in Section 3. The technical components that carry out each step of the process are described in detail in Section 4.

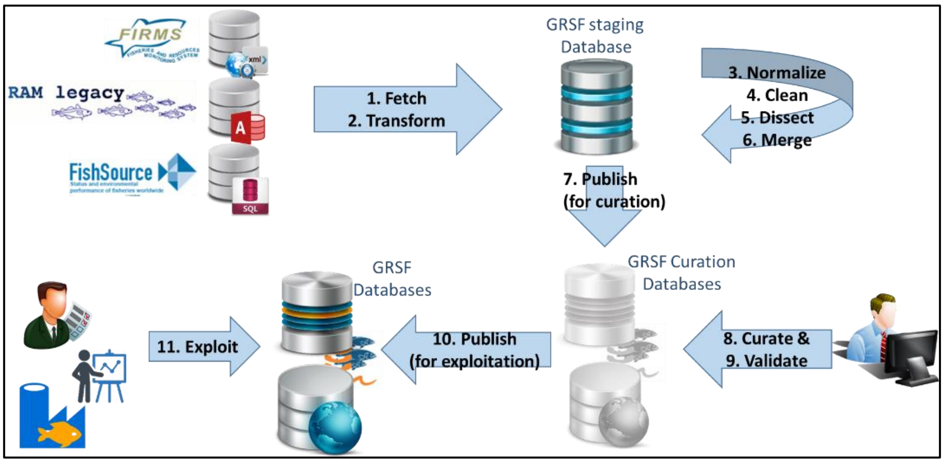

Fig. 2. The steps required for constructing and exploiting GRSF

Fetch. GRSF does not affect the data from the remote database sources. This means that the maintainers of the database sources will continue to update them in their own systems. For including the providers' data in the GRSF it is important to periodically fetch the raw data (in their original form) or the data in a different format or view if they are exposed using particular services (i.e. in other formats like JSON or XML). In particular FIRMS offers a set of services that exposes their contents in XML format, RAM publish their MS Access database in their website, and FishSource exposes specific parts of their relational database as JSON data through a set of services.

Transform. After fetching the data it is important to transform them so that they have a similar structure and semantics. At this stage data is transformed from XML, JSON and MS Access to RDF format. Specifically, data is transformed into instances of the MarineTLO ontology with respect to the identified GRSF requirements. Information harvested from the database sources will be mapped to the agreed GRSF standards, when not already compliant. Furthermore, during this step a set of proximity rules are applied (using the species, area and gear fields) for identifying similar records. This creates groupings of similar records that are being used in subsequent phases (during the curation & validation phase).

Normalize. The normalization step applies a set of normalization filters to the transformed data so that they are compliant with the GRSF standards. These filters may alter or add information related with a specific field to assist the instance matching functionalities of the merging process. Examples of normalization filters include the addition of 3alpha codes to a scientific name, the addition of ISO3 codes to a country or the specification of a water area standard regarding to a specific code. More information about the key normalization activities that are carried out are described in Section 3.

Clean. The data cleaning step includes all the necessary modifications of the source data in order to correct observed errors. The errors are being corrected either by the application of automatic filters (such us the genus capitalization for the scientific names) or by the notification of the sources to refine their data and re - harvest the altered content. More examples of data cleaning include the spelling correction in the scientific names, gear codes or water area codes.

Dissect. This step is important for complying with the standards, for traceability aspects. In some cases, sources contain aggregated information in their records. For example, in a single fishery record there could be included more than one species, fishing gears or flag states. These aggregated records are therefore dissected to produce new GRSF records, each containing one single value for the above-mentioned fields, and thus complying with the requirements for traceability.

Merge. This step ensures that the contents that have been added in the GRSF staging database are properly connected based on a set of criteria. This is achieved by linking records that have the same values on particular fields (specifically time-independent values) for producing a new single GRSF record. For example, if there are stock records having the same species and water area we can merge them into a single stock. During this process, we also use external knowledge to detect similarities among different names and terminologies used in the database sources (i.e. species names). The time-dependent information for the merged records will be kept distinct although collated and associated to the final merged GRSF record, with clear indication of the database source and the reference year.

Publish (for curation). The contents of the GRSF staging database are being replicated into a public GRSF database, which is actually a triple-store. The triple-store can be used as a reference endpoint for answering complex queries about stocks and fisheries records. Furthermore the contents are published in a data catalogue offered through the D4Science [5] infrastructure. These resources allow the experts inspecting the contents of the GRSF and curate them appropriately. During this step, Universally Unique Identifiers (UUID) and human readable semantic identifiers are generated and associated to each GRSF record. The former are generated based on a standard algorithm and are used to uniquely identify records. The latter are generated using various GRSF fields and populated with standard codes and allow the identification and interpretation of records by humans.

Curate & Validate. During this step, a community of experts browse over the GRSF records and curate them in various ways. At this stage, the GRSF records are in a pending status waiting for approval by a human expert. During this process, the experts are able to either approve or reject a record, as well as to suggest alternative processes for merging records and to attach annotations with a narrative text.

Publish (for exploitation). The GRSF records that has been approved during the previous phase are being published into public and read-only databases as final GRSF products that can be exploited from the communities of interest.

3 Data Normalization Activities

Two of the most important steps of the process, are the steps of data normalization and cleaning. During these steps we carry out several data normalization activities for guaranteeing that the results data are compliant with the set of GRSF standards. The GRSF standards has been created as the result of three technical working groups, with the participation of the owners of the GRSF data sources, the technical partners that are responsible for the construction and maintenance of the GRSF, as well as representatives of stocks and fisheries authorities from around the world. The standards describe several aspects of the information contained in GRSF from very basic ones (i.e. the proper capitalization of management entities names, use of international standards, etc.), to much more complex ones (i.e. identify similarities of records using various criteria). Below we describe these standards as well as the activities that were carried out for complying with them. We call the latter as Data Normalization activities.

3.1 Compliance with Standards

The key for interoperability is standardization, and GRSF is being constructed by exploiting international standards as much as possible. The use of these standards has been agreed in two dedicated technical working group meetings, with the participation of representatives of the used data sources. At this point, we should also describe that standards are being partially used for the underlying sources as well (i.e. FIRMS uses 3-Alpha codes for identifying marine species, while other source do not). Below we describe in detail the standard schemes that are exploited in GRSF.

Marine species. There are various ways for identifying a marine species; usually we use their common names (e.g. yellowfin tuna), however it is not the best alternative since there are multiple common names (with values in several different languages and multiple names used even for single countries). One alternative for identifying species is their scientific name (or binomial name) which is composed of two parts, the first being the genus name and the second is the specific species name (e.g. Thunnus albacares). Another alternative for identifying species is their 3Alpha code. 3Alpha codes have been introduced by ASFIS[9], and consist of three letters that uniquely identify the species. In most of the cases, the codes have been derived either from the scientific name of the species, or by their common name in English (e.g. YFT is the 3Alpha code for yellowfin tuna). In all other cases, the three letters are assigned at random. In the absence of 3Alpha codes GRSF has adopted the aphia ID[10] as an alternative standard.

Water areas. Similarly to marine species, water areas can have commonly used names. However they are not adequate for identifying the area itself, since the boundaries of the area are not clearly defined. A more accurate method is to describe them

[9] http://www.fao.org/fishery/collection/asfis

[10] http://www.marinespecies.org/aphia.php?p=webservice

using polygons that are formulated using pairs of geographic coordinates. A polygon is an accurate description of an area, since it can take any shape. A simpler abstraction is to use bounding boxes for modeling a water area. Compared to the polygons the bounding boxes are less detailed, however it is much simpler to perform geographic calculations using them (i.e. find overlapping or adjacent areas). Apart from the above there is also a coding system[11] from FAO that allows identifying water areas using codes (e.g. the aegean sea has the FAO water area code 37.3.1). The FAO area codes are the primary standard used by GRSF but eligible standards are also the ICCAT[12], Pacific Tuna, RFB[13] competence areas and GFCM[14] codes.

Countries/States. Countries can be described using their ISO 3166-1 Alpha-3 codes. These codes are composed of three letters and represent countries, dependent territories and special areas of geographical interest (e.g. the ISO Alpha-3 code for Greece is GRC).

Fishing Gear. The Coordinating Working Party on fishery statistics (CWP)[15] provides a mechanism to coordinate fishery statistical programmes of regional fishery bodies and other inter-governmental organizations with a remit of fishery statistics. CWP adopted in 1980 a labeling and classification standard for fishing gears [6] that led to the creation of the International Standard Statistical Classification of Fishing Gears (ISSCFG). The standard assigns an acronym and a classification code that can be used for identifying gears of the same type. For example portable lift nets are identified using the acronym LNP, while boat-operated lift nets are identified using the acronym LNB. The former has the classification code “05.1.0” while the latter has the code “05.2.0”. The common prefix of the classification codes (e.g. “05”) allow us identifying that they are similar types of fishing gears, in this case lift nets. The most recent revision of the standard has been carried out in 2016 and contains new classification codes fishing gears.

3.2 Identification of Unique Records

A crucial step for the proper integration of stocks and fisheries data from heterogeneous sources is the identification of unique records (single records that are co-references in different sources). The identification of a single record will allow carrying out merging activities in the sequel. For example, this would allow merging the time-dependent information of records coming from different sources and deliver a single GRSF record. To this end, it is important to define which are those fields that make a record unique.

The GRSF standard methodology defines the uniqueness of a stock record using: (a) the fish species it contains and (b) the water area it occupies. As regards fisheries their uniqueness is defined using: (a) the fish species it contains, (b) the water area it

[11] http://www.fao.org/fishery/area/search/en

[12] https://www.iccat.int/en/

[13] http://www.fao.org/fishery/rfb/collection/en

[14] http://www.fao.org/gfcm/en/

[15] http://www.fao.org/fishery/cwp

occupies, (c) the management entity that operates the fishery, (d) the flag state under which the fishery is operated, and (e) the fishing gear that has been used.

To avoid potential disambiguation and naming issues, since different source could use different names for their resources (i.e. names marine species, water areas, fishing gears, etc.) which could result in errors, the identification of records is carried out after compliance with standards activity that was described before.

3.3 Semantic Identifiers

In addition to the compliance with international standards and the identification of unique records, it has been decided to construct global identifiers for GRSF records that are human readable. These identifiers are called semantic identifiers in the sense that their values allow identifying several aspects of a record. The identifier is a concatenation of a set of predefined fields of the record in a particular form. The rationale is that users will be able to recognize important information about a stock, just by inspecting the semantic identifier. To keep the length of the identifier in a reasonable number, it has been decided to use the standard values or abbreviations where applicable. For example consider the following semantic identifier of a stock "`ASFIS:lub+FAO:51.6`" that denotes the stock record is about the species "`lub`" (with respect to the 3Alpha code of the ASFIS system) and the water area with code "`51.6`" (with respect to the FAO coding system for areas).

The fields of the identifier following the pattern `<SYSTEM:CODE>`. The first field denotes the classification system that was used and the second is the actual code. In addition the fields are concatenated using the character '+' as a separator, and the fields are reported in particular order. If there are more than one values for a particular fields in the record then they are all reported, using the same pattern and they are concatenated using the character ';'. It is evident from the above example that for stock records the first field is the species and the second one is the water area of the record. For the case of fishery records the semantic identifier contains the following fields (in the given order): (1) species, (2) water areas, (3) management entity, (4) jurisdiction area (5) flag state and (6) fishing gear. An indicative semantic identifier of a fishery record is "`asfis:COD+fao:21.3.M;rfb:NAFO+grsf-org:INT:NAFO+rfb:NAFO+iso3:LTU+isscfg:03.1.2`". Notice that in this fishery records that are two different water areas (second field) described in the semantic identifier. Finally, if for a field there is not any information in the record then an empty string is added for that field.

3.4 Multiple Values Prioritization

When we integrate data from heterogeneous sources, it is inevitable that we might end up with multiple values about a particular aspect of the same resource, each one coming from a different resource. This is also true for the case of GRSF, and it becomes an issue when there are multiple values for the time-independent information of a record. An indicative example is the name of a record (either stock or fishery); if a

GRSF record is the result of merging of 3 original records (from the corresponding data sources) then we will end up with 3 different names of the record. This is usual, since the original data sources use their own policies for naming their records.

In order to resolve this issue it has been decided to adopt a prioritization policy for multiple values. This means that we prioritize the sources and whenever there are such situations, we will use the value coming from the top source. If the top source does not contribute with a value in the record then we move to the next source in the order and so on. Particularly for GRSF, we prioritize values about the names of the records and the assessment areas with the following order: (1) FIRMS (2) FishSource, (3) RAM.

3.5 Records Similarities

Apart from being a global registry, GRSF aims at supporting the experts with the stock and fishery assessment activities. Part of these activities is the identification of similar records that could potentially be merged to single records and produce new knowledge. To this end, during the merging step we carry out several comparisons between records in order to identify similarities between records. The following table shows the criteria that should apply for considering two records similar. For example if two records have species that have the same genus and appear on adjacent areas then they are considered as similar. The criteria are applied for fishery records as well, with the amendment that apart from the criteria shown in the table the records under comparison should appear under the same group of fishing gear with respect to the fishing gears hierarchy.

Table 1. Criteria for defining similarities between records

	Area		
	Same	**Adjacent**	**Overlapping**
Species	✓ (fisheries only)	✓	✓
Genus	✓	✓	✓

For this reason, we first identified the adjacent and overlapping areas of the records. We used the bounding boxes that represent the geographical coverage of the of the records as they have been derived from the original sources and used an R script[16] for defining if they are the same, adjacent or overlapping. Although, RAM data source did not contain any information about the bounding boxes for each record, it was using a name for the area of each record and a bounding box for the area, and this is what we have used for the comparisons. Of course, this means that a record occupied the entire area, instead of a smaller region and this could raise issues. However this is not a problem since we are proposing similarities that will be validated from experts in subsequent phases.

[16] https://www.r-project.org/

3.6 Data Cleaning

In order to maximize the quality of GRSF we supported data cleaning activities so that the textual information appears in a common and uniform way and observed errors are being corrected as much as possible. Below we describe some of the fields that were cleaned, as well the activities carried out.

- Scientific Names of species: whenever the scientific names of the species existed, we ensured that the first character of the genus was always a capital letter and the rest of it as well as the specific epithet use letters in lower case (i.e. Thunnus albacares).
- Management Entity: we used capital letters for the first characters of the terms of the management entity and also constructed an abbreviated acronym from the capital letters (i.e. Northwest Atlantic Fisheries Organization – NAFO).
- Water areas: the FAO water areas codes in the sources may include extra points or zeros, which must be, eliminated (i.e. FAO area 05 → FAO area 5).
- Gears: the gear ISSCFG codes may also contain extra points or zeros, which must be, eliminated (i.e. 01.2.0 → 01.2).
- Others: other possible errors that have not been predicted are reported to the maintainers of the data sources. They refined their data and the altered content was then harvested and imported in GRSF.

4 Software Components and Architecture

The D4Science infrastructure and gCube technology [5,7] enable the development of Virtual Research Environments (VREs) that provide the users with a web-based set of facilities to accomplish various tasks. For the purpose of GRSF, we developed the appropriate VREs acting as a gateway for the “one stop shop” for stocks and fisheries records. More specifically we exploit the data cataloguing facilities of the infrastructure for manipulating and exposing GRSF records to the wide audience.

The core component for constructing GRSF is MatWare [8]. MatWare is a framework that automates the process of constructing semantic warehouses. By using the term semantic warehouse we 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 (e.g. SPARQL endpoints, HTTP accessible files, JDBC connections, several file format transformers). The fetched data are transformed into RDF descriptions using appropriate mappings [9], and stored in a RDF triplestore supporting several levels of description for preserving provenance information. One of its distinctive features, is that it allows evaluating the connectivity of the semantic warehouse. Connectivity refers to 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, while making evident how each source contributes by using a set of connectivity metrics. MatWare is a fully configurable tool and can be easily extended using plugins. For the purposes of GRSF construction we have extended it with plugins for transforming the data from their original formats, plugins for supporting the merging and dissection steps, as well plugins for publishing the data into the catalogue supporting both the curation and validation phase, as well as the consumption phase.

Fig. 3 shows the overall technical deployment for the construction and maintenance of the GRSF. MatWare is responsible for the activities that construct the GRSF (as they are described in Section 2.4) and publishing them in the GRSF Knowledge (GRSF KB) and in the GRSF Catalogue. For the latter it exploits the component Data Catalogue publisher which carries out the necessary activities for ingesting GRSF records into the CKAN-based Catalogue instance offered by the D4Science infrastructure. Finally all the above components are controlled and interacted through the D4Science portal facilities of the GRSF VREs.

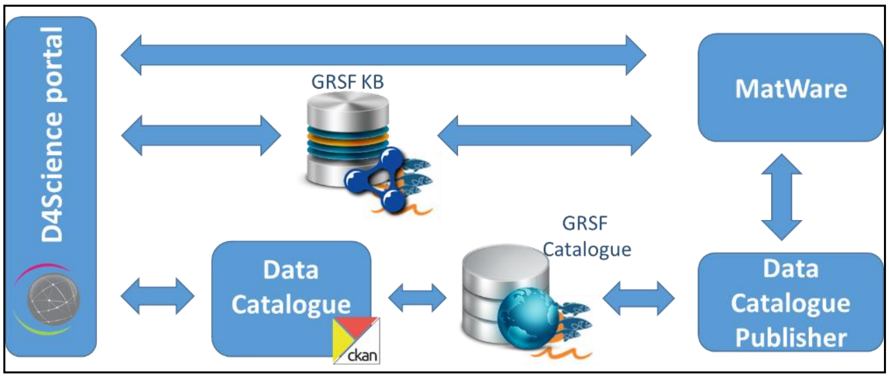

Fig. 3. The GRSF construction deployment setting

5 Conclusion – Future Work

The collation of information for the monitoring of fish stocks and fisheries is a difficult and time-consuming task, as the information is scattered across different databases and modelled using different formats and semantics. We introduced a process for providing a unified view of several stock and fisheries databases, by relying on semantic web technologies and innovative hybrid data infrastructures. The resulting Global Record of Stocks and Fisheries integrates data from three data sources, and contains more than 9,500 records about stocks and fisheries. It can be seen as a core knowledge base supporting the collaborative production and maintenance of a comprehensive and transparent global reference set of stocks and fisheries records. This is accomplished because of the processes that were applied during the construction, that guarantee the unique identification of stock and fisheries and the easy access to all the available information associated to a particular stock or fishery. In addition, during the validation step, the experts can validate the information of the GRSF records which also allows them spotting errors in their original sources, because their provenance is also preserved.

In order to maximize the quality of the GRSF contents, as well as their potential exploitation, we carry out a set of data normalization activities during the dissection and merging steps. These activities assert that the records and their accompanying information is valid and it is compliant with international standards where this is feasible. Table 2 summarize some statistics about GRSF.

Table 2. Summary of the information fetched and integrated into GRSF

	FIRMS	RAM	Fish-Source	GRSF
Stock Records	866	1294	1156	2,918
Fishery Records	271	-	3,112	8,719
Species	612	349	488	1, 494
Water Areas	275	803	418	1,496
Fishing Gears	33	-	50	83
Timeseries	9,242	226,725	47,656	283,623
Similar Records	-	-	-	18,524

Some activities that are worth further work and research (a) investigation of whether machine-learning techniques could be exploited for automating or assisting the curation and validation of GRSF records, and (b) exploitation of advanced discovery services based on spatio-temporal information.

Acknowledgments. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the BlueBRIDGE project (Grant agreement No 675680).

References

1. R. Hilborn and C. Walters (Eds.). Quantitative fisheries stock assessment: Choice, Dynamics and Uncertainty. Springer Science & Business Media, 2013, ISBN 978-1-4615-3598-0.
2. Y. Tzitzikas, Y. Marketakis, N. Minadakis, M. Mountantonakis, L. Candela, F. Mangiacrapa, P. Pagano, C. Perciante, D. Castelli, M. Taconet, A. Gentile, G. Gorelli. Towards a Global Record of Stocks and Fisheries. In Proceedings of the

8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2017), Chania, Crete Island, Greece, September 21-24, 2017.

3. Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos, L. Candela. Unifying Heterogeneous and Distributed Information about Marine Species through the Top Level Ontology MarineTLO. Emerald Group Publishing Limited 50(1) 2016. http://dx.doi.org/10.1108/PROG-10-2014-0072
4. M. Pham, S. Alse, C.A. Knoblock, P. Szekely. Semantic Labeling: A domain-independent approach. In procs of the 15th Int. Semantic Web Conference (ISWC 2016), Japan. 2016
5. L. Candela, D. Castelli, A. Manzi, P. Pagano. Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. Int.Symposium on Grids and Clouds (ISGC) 2014, Proceedings of Science PoS (ISGC2014).
6. C. Nédélec, J. Prado. Definition and classification of fishing gear categories. Définition et classification des categories d’engins de péche. Definición y clasificación de las diversas categorías de artes de pesca. FAO Fisheries Technical Paper 222 (1990).
7. M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, F. Sinibaldi (2018) The gCube system: Delivering Virtual Research Environments as-a-Service. Future Generation Computer Systems doi: 10.1016/j.future.2018.10.035.
8. Y. Tzitzikas, N. Minadakis, Y. Marketakis, P. Fafalios, C. Allocca, M. Mountantonakis, I. Zidianaki. MatWare: Constructing and Exploiting Domain Specific Warehouses by Aggregating Semantic Data. In procs of the 11th Extended Semantic Web Conference (ESWC'14), Anissaras, Crete, Greece, May 2014.
9. Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, M. Doerr. X3ML Mapping Framework for Information Integration in Cultural Heritage and beyond. Int. Journal on Digital Libraries, pp 1-19. Springer. DOI 10.1007/s00799-016-0179-1.