NeOn-project.org

NeOn: Lifecycle Support for Networked Ontologies

Integrated Project (IST-2005-027595)

Priority: IST-2004-2.4.7 – “Semantic-based knowledge and content systems”

D7.2.3 Initial Network of Fisheries Ontologies

Deliverable Co-ordinator: Caterina Caracciolo (FAO)

Author: Caterina Caracciolo (FAO)

Contributor: Juan Heguiabehere (FAO), Valentina Presutti (CNR), Aldo Gangemi (CNR)

Deliverable Co-ordinating Institution:

Food and Agriculture Organization of the United Nations (FAO)

Document Identifier:	NEON/2009/D7.2.3/v1.1	Date due:	February 28, 2009
Class Deliverable:	NEON EU-IST-2005-027595	Submission date:	September 15, 2009
Project start date:	March 1, 2006	Version:	1.1
Project duration:	4 years	State:	Final
		Distribution:	Public

NeOn Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Community; grant number IST-2005-027595. The following partners are involved in the project:

Open University (OU) – Coordinator Knowledge Media Institute – KMi Berrill Building, Walton Hall Milton Keynes, MK7 6AA United Kingdom Contact person: Martin Dzbor, Enrico Motta E-mail address: {m.dzbor, e.motta} @open.ac.uk	**Universität Karlsruhe – TH (UKARL)** Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB Englerstrasse 28 D-76128 Karlsruhe, Germany Contact person: Peter Haase E-mail address: pha@aifb.uni-karlsruhe.de
Universidad Politécnica de Madrid (UPM) Campus de Montegancedo 28660 Boadilla del Monte Spain Contact person: Asunción Gómez Pérez E-mail address: asun@fi.upm.es	**Software AG (SAG)** Uhlandstrasse 12 64297 Darmstadt Germany Contact person: Walter Waterfeld E-mail address: walter.waterfeld@softwareag.com
Intelligent Software Components S.A. (ISOCO) Calle de Pedro de Valdivia 10 28006 Madrid Spain Contact person: Jesús Contreras E-mail address: jcontreras@isoco.com	**Institut 'Jožef Stefan' (JSI)** Jamova 39 SI-1000 Ljubljana Slovenia Contact person: Marko Grobelnik E-mail address: marko.grobelnik@ijs.si
Institut National de Recherche en Informatique et en Automatique (INRIA) ZIRST – 655 avenue de l'Europe Montbonnot Saint Martin 38334 Saint-Ismier France Contact person: Jérôme Euzenat E-mail address: jerome.euzenat@inrialpes.fr	**University of Sheffield (USFD)** Dept. of Computer Science Regent Court 211 Portobello street S14DP Sheffield United Kingdom Contact person: Hamish Cunningham E-mail address: hamish@dcs.shef.ac.uk
Universität Koblenz-Landau (UKO-LD) Universitätsstrasse 1 56070 Koblenz Germany Contact person: Steffen Staab E-mail address: staab@uni-koblenz.de	**Consiglio Nazionale delle Ricerche (CNR)** Institute of cognitive sciences and technologies Via S. Martino della Battaglia, 44 - 00185 Roma-Lazio, Italy Contact person: Aldo Gangemi E-mail address: aldo.gangemi@istc.cnr.it
Ontoprise GmbH. (ONTO) Amalienbadstr. 36 (Raumfabrik 29) 76227 Karlsruhe Germany Contact person: Jürgen Angele E-mail address: angele@ontoprise.de	**Food and Agriculture Organization of the United Nations (FAO)** Viale delle Terme di Caracalla 1 00153 Rome Italy Contact person: Margherita Sini E-mail address: margherita.sini@fao.org
Atos Origin S.A. (ATOS) Calle de Albarracín, 25 28037 Madrid Spain Contact person: Tomás Pariente Lobo E-mail address: tomas.parientelobo@atosorigin.com	**Laboratorios KIN, S.A. (KIN)** C/Ciudad de Granada, 123 08018 Barcelona Spain Contact person: Antonio López E-mail address: alopez@kin.es

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed to writing parts of this document:

INRIA, UKARL, USFD.

Change Log

Version	Date	Amended by	Changes
0.1	15/11/2008	Caterina Caracciolo	First Draft
0.2	09/12/2008	Caterina Caracciolo	Second draft
0.3	10/02/2009	Caterina Caracciolo	Reorganization of contents
0.4	05/03/2009	Caterina Caracciolo	Added section on revision/validation of ontologies
0.5	10/03/2009	Caterina Caracciolo	Updated Introduction and Discussion. First draft conclusions.
0.6	15/03/2009	Caterina Caracciolo	Sec. 3 revised. Version Sent to QA
0.7	20/02/2009	Caterina Caracciolo	Accommodated comments from QA. Revised section on discussion and conclusions.
0.8	30/03/2009	Andrew Galea Debono	English revision.
0.9	09/04/2009	Aldo Gangemi	Added ASFA-related chapter, ASFA references in existing sections, and added/revised bibliographic references
1.0	11/04/2009	Caterina Caracciolo	Harmonization of chapter on ASFA with the rest of the document. Sections on ASFA reengineering, mapping to RTMS and modularization moved into Chapter 3 (Sec. 3.3). Sections on alignment of SKOS and LMM, example from asfad.owl moved to Annex. Added motivations for ASFA reengineering and inclusion in the network. Future work for ASFA included in general plan for future work. Uniform style to bib/ref applied. Added missing references and cross-references. Updated Executive summary.
1.1	01/09/2009	Caterina Caracciolo	Accommodated requests by EU reviewers: added Annex VII....

Executive Summary

This document describes and discusses the fisheries ontologies developed for use within the Fish Stock Depletion Assessment System (FSDAS). All ontologies are publicly available from the FAO website, from http://www.fao.org/aims/neon.jsp. This document is organized as follows:

In Chapter 1 we introduce our work and in Chapter 2, we introduce the topics in fisheries covered by the network of ontologies presented in this deliverable.

In Chapter 3, we present the first network of fisheries ontologies. In this chapter we provide a detailed description of the RTMS data used; an account for the validation and revision process that followed the previous release of (individual) ontologies; an account of the work on reengineering of the ASFA thesaurus together with a sample of links to include ASFA in the network of RTMS-based ontologies and a discussion about it; an account of the process followed when reengineering the RTMS data as a network of ontologies; and a summary of the features of the network. We have been able to produce a network based on linking information stored either in the RTMS, or in other textual corpora (i.e. fisheries fact sheets). The ASFA ontology reengineered from the ASFA thesaurus is currently not, or only marginally, included in the network, as only a small sample of links is provided. However, it offers the basis to understand the issues related to linking ontologies adopting different modelling styles. Further work will be devoted to automatic and semi-automatic linking (alignments) of these ontologies, especially for what concerns the linking at instance level.

In Chapter 0, we discuss the issues which arose during our work. In particular, we highlight the modelling decisions adopted in the case of the RTMS-based ontologies, and report on the technical limitations found when implementing the network with the tools currently available (detailed feedback was sent to the developers whilst our work was carried out).

Our conclusion, reported in Chapter 5, is that the reengineering of resources such as the reference data for time series is possible, although the tools available need to be improved in order to make the process smooth and therefore more widely adopted. Future work includes experimenting with methods and tools to connect ontologies in a network on the basis of information available from a variety of sources.

This document also includes seven Annexes: the list of naming conventions adopted (Annex I), an essential glossary of fisheries terms (Annex II), a list of acronyms (Annex III), the list of correspondences between RTMS tables and the ontologies produced (Annex IV), the hierarchy of meta codes used in RTMS (Annex V), and details about the reengineering of ASFA (Annex VI). Finally, an annex (Annex VII) was added upon request of the EU reviewers, to clarify the methodologies and plugins used, to highlight the position of the NTK with respect to competitors, and assess how well NTK achieves the goal of supporting the entire lifecycle of networked ontologies.

Table of Contents

List of tables

List of figures

1 Introduction

The WP7 case study is concerned with the creation of an ontology-driven Fisheries Stock Depletion Assessment System (FSDAS). Such a system is meant to exploit a number of FAO datasets on the fisheries domain, some of which are reengineered as a network of ontologies. In this deliverable we present an initial network of ontologies of fisheries based on reference data for statistical time series. We also present the reengineered version of the ASFA Thesaurus[1] as an ontology, together with some preliminary work and a discussion about its inclusion in the network. This deliverable does not include the "application ontologies" created for FSDAS, as those ontologies are application-oriented and will be included in D7.6.2 [D7.6.2].

"A network of ontologies is a collection of ontologies related to each other via a variety of different relationships such as mapping, modularization, version, and dependency relationships. We call the elements of this collection Networked Ontologies" (cf. NeOn deliverable [D1.1.5][2] forthcoming). On the one hand, when the relationship requires the existence of correspondences between ontologies (and thus overlap in their modelled domain), it is called mapping. On the other hand, when the relationship involves two disjointed ontologies, it is called e-connection (however, throughout this document we prefer to use the more general term "link"). This latter case (in which the relationships can be conveniently conceived as super-roles relating the two domains and therefore be used as constructors for complex concepts) is largely used in the network we present here. Other types of relationship, such as version, modules and dependency (cf. [D1.1.5]), are not used in this network.

This deliverable builds on Deliverable D7.2.2 [D7.2.2], in which six ontologies[3] were produced on the basis of the reference data for statistical time series used by the Fisheries and Aquatic Information and Statistical Service (FIES) department of FAO. Reference data is organized into tables, called reference tables (stored as a relational database managed by a RDBMS) that we refer to as RTMS.[4] The six ontologies included in D7.2.2 are now improved in several ways (according to feedback gathered from different sources, cf. Sec. 3.2) and linked in a network. The work performed in D7.2.2 served as an exploration of the possibilities and risks of the reengineering of FAO data into ontologies. The result of that experiment was all-in-all positive, as we found that the reengineering is possible, although technically not always straightforward, and we learned useful lessons. From the modelling point of view, we found that some apparently intuitive notions of what we consider as class and instances in the fisheries domain had to be revised and adapted to the reality of the data at hand (see Sec. 3.2 and Sec. 4.1.2). Moreover, on the basis of what we learned we could provide useful feedback to FIES on the organization of the data. From the implementation point of view, we found some difficulties that were reported in Sec. 8 in [D7.2.2]. Now some of those difficulties have been overcome, while others are currently under consideration. In particular, we are now better off with run-time access to DB according to

[1] http://www4.fao.org/asfa/asfa.htm

[2] The model presented in D1.1.5 is particularly aimed at supporting ontologies in OWL2 and F-logic. However, the fact that the ontologies presented here are all in OWL 1.0 is not a problem, as the language is backward compatible. We preferred to stick to OWL 1.0 because at the time of this work, OWL2 had not yet reached the status of w3c recommendation.

[3] About: land areas, FAO water divisions, biological entities, fisheries commodities, gear types, vessel types. All ontologies are available at: http://www.fao.org/aims/neon.jsp

[4] The entire machinery managing the reference tables is called RTMS (Reference Tables Management System); the same acronym is also used to refer to the entire set of reference data. In the course of this document, we also use "FIGIS", as the actual name of the database storing the reference data – not to be confused with FIGIS, the FAO project on fisheries.

ontological models, and special attention is being devoted to the linking of ontologies through instances.

There are many reasons for reengineering the RTMS. From an information management perspective, the RTMS does not manage the entire data lifecycle and, for this task, a variety of systems are in use, each for specific purposes and specific users. Therefore, one expected advantage of moving to ontologies and using the NeOn technologies is that the entire lifecycle could be supported in a coherent way: ontology design (implementation, methodologies), editorial workflow, connection to other sources of information, mapping generation. In particular, the complexity of the RTMS database schema, together with the fact that data comes from a variety of sources (e.g. international organizations, regional bodies, governments, cf. Chapter 2) makes the maintenance of the data a complex task, which could be greatly simplified if the graphical interfaces and the user-interaction model were smoothed. Also, the use of standards for both the domain modelling and data encoding is expected to facilitate data reuse and sharing. This would also improve the possibility of using off-the-shelf tools.

From a functional point of view, reference data is used to index time series on fisheries collected by FIES[5] (cf. Sec. 2.1). Therefore, it is ultimately used to answer questions (input as a query through an online query panel) such as: “what was the quantity of frozen crustacean imported by the Russian Federation in 1997?”[6], or “what is the total amount of capture production of Japan in South West Pacific?”[7]. The ontologies reengineered from the RTMS are expected to continue to provide this functionality (granted that the machinery involved is opportunely adapted). However, ontologies connected through a network are expected to serve as a basis to provide functionalities currently not supported by the RTMS organization of data. For example:

1. RTMS does not provide any form of constraints of data and it is possible to compose queries that have actually no meaning in the real world (e.g. “catch of yellowfin tuna (*Thunnus Albacares*) in the Mediterranean Sea”, while yellowfin tuna is found in open waters of tropical and subtropical seas worldwide).

2. RTMS provides little information useful to integrate missing data. For example, there is no explicit notion of proximity between countries or water areas. This information can be extremely useful, as when data about one area (land or water) is missing, it is often useful to look at neighbouring areas, or at regions sharing some specific features (e.g. climatic zones, shore on the same sea, contiguous or non-contiguous water areas where a given species can be found).

Finally, what is perhaps the main limitation of the RTMS as a database is that its complex design (cf. Sec. 3.1) makes data maintenance rather cumbersome (especially data input and validity checks). However, that complex design also offers the advantage of generality and, consequently, flexibility, allowing for a high degree of modularity and extensibility. The natural tension between these two dimensions - complexity and generality - could be overcome by adopting an ontological layer through which to access the data. In fact, data modelled according to ontologies can be more flexibly reorganized and exploited for different needs. Regarding the actual relationships between the ontological layer and the database, we can identify at least three scenarios. Before sketching these scenarios, we stress the fact that the importance of data maintenance cannot be overestimated, and any scenario will have to be assessed against its advantages in terms of data maintenance over the other options. A short description of the three scenarios envisaged follows, together with a discussion of their pros and cons:

[5] For a list of statistical databases maintained by FAO see: http://www.fao.org/fishery/statistics/en.

[6] The database involved is: “Commodities 1976-2006”.

[7] The database involved is: “Global capture production”.

1. *Ontologies are used to access and visualize data residing in the DB.* Data stays in the DB and the ontologies are used to access and visualize data at run-time, as well as to expose data to both human users and applications. The connection between the DB and the ontologies is a one-way connection, and data is maintained in the database.

2. *Ontologies are used to store, access and visualize data residing in the DB.* Ontologies are then used not only for accessing and exploiting data, but also to store it in the DB. This scenario features a two-way connection with the database, so that data is maintained through the ontologies and the technicalities of database storage are kept away from the user.

3. *Data is migrated to ontologies (and RDF data sets).* In this scenario, the relational database fades out and data is maintained not only “through” but also “in” the ontologies. The (network of) ontologies here presented would then serve as an intermediate step to migrate data from the relational database to RDF.

The first scenario is the one embodied in this deliverable (data resides in the DB, and it is exposed to human users and applications through ontologies). In some ways, it is the most conservative one, as the DB is kept almost as it is and the ontologies are simply added to the RTMS to enhance some of its functionalities. The first advantage of this scenario is that the efficiency of RDBMS remains available for exploitation; secondly, it spares the expensive and error prone processes of data conversion; thirdly, all applications that access the data will continue to work. The disadvantage is that data maintenance would be done at the level of the DB, which would bring us back to a situation of a non-continuous lifecycle. As a consequence, the setting to access the data may have to be regenerated every time the DB is updated. Moreover, a possible drawback is that, in some cases, modifications to the DB schema may be required, by means of views. This is not ideal, because of the effort involved and because of different levels of ownership on the data.

The second scenario consists in a situation where data resides in the DB, but ontologies serve as an intermediate layer to store, retrieve and expose data to applications. This option implies that a complete round-trip between ontology and DB is possible, with resulting advantage in terms of data maintenance. Moreover, the data could actually be physically kept in the DB, with the advantage already mentioned (in point 1) that third applications accessing the data may continue to work. This scenario is extremely interesting, because it allows for full exploitation of the advantage of semantically-driven technologies without running the risk of disrupting current functionalities and applications of the data. Moreover, it could be taken as a short or medium term solution, to allow fisheries experts to evaluate the technologies and make sure that no applications are harmed by the switch of technology.

Finally, the third scenario (all data is moved to ontologies, encoded in RDF) is the most drastically different from the current situation. The advantage of this scenario is that it would be a complete migration to semantically-oriented technologies with consequent full exploitation of all their potentialities. Moreover, reducing all data to RDF triples would make the data set extremely flexible. However, this migration would be a major undertaking with expected “unforeseen” consequences due to the many third applications using the RTMS either directly or indirectly. Apart from its technical pros and cons, various possible sources of risks should be taken into consideration when pondering on this solution. First of all, the ontologies produced for this deliverable are arguably the most important in the RTMS, but also certainly a relatively small portion of the data it contains. Therefore, a deep analysis of all the fragments should be made in order to understand the methods of conversion for all of them. All dependencies should be checked, between data. The machinery that uses the reference data to query the time series should also be converted appropriately.

Currently, we have modelled as ontologies the domains covered by the reference data, while the actual data is in the database. Data is then maintained in the DB and also accessed straight from there in the case of all other applications using the RTMS. Note that, in this way, the network of ontologies presented in the current deliverable did not go through the entire lifecycle: ontologies went from the first step of iteration of conceptualization and population, directly to the publishing in the production environment. The intermediate of validation of data and update was skipped because tools are not yet mature to take this big step.

Independently of the scenario that will be implemented, a number of aspects need to be taken into consideration while performing the reengineering of the RTMS. In the following, we list those we considered during our work.

1. **The modelling adopted should:**
    - **a.** **reflect the "nature" of the domain described**;
    - b. **comply with good modelling style, methodologies and practice**;
    - c. **be feasible,** i.e. it should allow access to the desired data in the database: this depends on the functionalities of the tools available for that, and on the underlying implementation of the database;
    - d. **allow for backward compatibility with current applications**, i.e. the accessed data should be available to these applications or, at least, it should minimize the effort required to adapt them.
2. **The reengineering process should be:**
    - **a.** **repeatable,** in case data is updated, or mistakes are found. The process should be easily repeatable even if it is expected to apply it to migrate data into the new format;
    - **b.** **simple,** otherwise it will not be actually implemented.

As for the data linking ontologies (and so forming the network), it was extracted either from the database or from other sources, mainly FAO fact sheets on fisheries. When data was extracted from other sources than RTMS, our approach was to also store that data in a DB and access it through the tool we used for data reengineering (cf. Sec. 3.3).

The output of our work on RTMS reengineering is the following:

1. the present document is meant to a) contain all the information needed for a non-FAO user to understand the network of ontologies; b) trace the modelling and implementation decisions made during our work; c) report issues and problems encountered, so as to serve as reference for future work and allow their solution at a later time.
2. a network consisting of ontologies (T-box) about: taxonomic classification of biological entities, ISSCFC and HS classification of fisheries commodities, ISSCAAP classification of species (with links to taxonomic classification of biological entities and ISSCFC commodities), FAO division of water areas, large marine ecosystems (with links to FAO divisions), geopolitical ontology, exclusive economic zones (with links to geopolitical ontology), ISSCFG classification of gear types, ISSCFV classification of vessel type and size, stocks (with links to taxonomic classification of biological entities and FAO divisions). All ontology schemas (T-box) are made available to the public through the FAO website: http://www.fao.org/aims/neon.jsp. All ontologies are endowed with comments so as to make their exploitation possible also independently of this document.
3. the instances (A-box) corresponding to the ontology schemas, also publicly available at the same website, so as to allow users to exploit the actual data associated to the network of ontologies, independently of the connection to the RTMS database.

The output of our work on ASFA is an ontological version of the thesaurus (cf. Sec. 3.3), and a sample set of mappings between the reengineered ASFA and the RTMS-based ontologies (Sec. 3.3.5). Some sample modules induced by the mappings are also detailed in Annex VI.

The rest of this deliverable is organized in the following way. In Chapter 2 we introduce the fisheries domain. In Chapter 3 we describe in detail the RTMS, the validation and revision phase that followed the release of the ontologies in D7.2.2, the reengineering of ASFA, a discussion about the inclusion of the reengineered ASFA into the network of RTMS-based ontologies, and the features of the current network. In Chapter 0 we highlight and discuss issues that may be of interest for future work, while in Chapter 5 we draw conclusions and hint at future work.

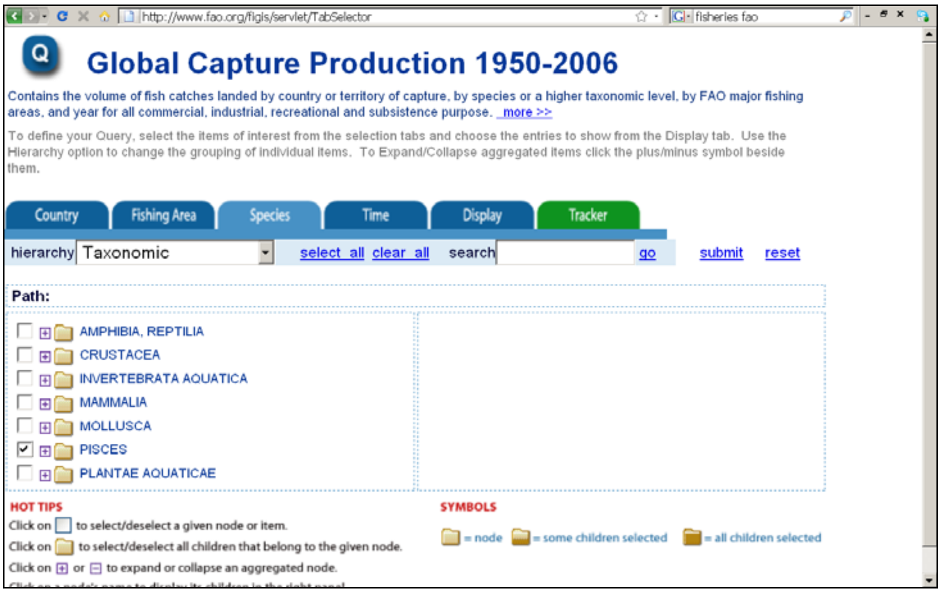
The naming conventions adopted in the making of the ontologies are described in Annex I. A glossary of relevant concepts is included in Annex II, and a list of acronyms is included in Annex III. In Annex IV and V we further describe the database and the correspondences between the tables in the database and the corresponding ontologies. Finally, in Annex VI we provide more details about the reengineering of ASFA and details about a sample modularization of ASFA based on the links with RTMS.

2 The fisheries domain

In this chapter we introduce the concepts that are modelled in the fisheries ontologies we produced. Our aim is to provide the reader with a very general background on fisheries, with special attention to the notion necessary to understand fisheries metadata for time series. We also provide some details about the data collected by FIES. With this background, the reader should be able to understand the rationale of the ontologies included in the network, and the meaning of their links. First of all, we introduce the reference data and the applications using it (Sections 2.1 and 2.2). Then, we introduce the specific topics covered by the fisheries network of ontologies: biological entities (Sec. 2.3), fishing areas (Sec. 2.4), land areas (Sec. 2.5), the concept of fish stock (Sec. 2.6), fisheries commodities (Sec. 2.7), fishery fleet (Sec. 2.8), vessel types (Sec. 2.9), and gear types (Sec. 2.10).

2.1 Reference data, coding systems and real world entities

The FAO Fisheries and Aquaculture Information and Statistics Service (FIES) collates a large amount of data concerning several aspects of fisheries. The data collected includes statistical time series, reports, maps and GIS data, fact sheets (about species, fishing techniques, stocks, and other related topics), for which metadata is fundamental. An important part of this metadata is used for the indexing of statistical time series. A time series is a sequence of statistical observations which are ordered in time and/or space. FIES collects observations about captures, aquaculture production, catches, fleets, trade of commodities, and consumption[8] [FISTAT]. Any piece of statistical data is referenced by the following dimensions: time (in years), space (land and/or water areas), and the variable representing the observed object (e.g. biological species). For example, we can have statistical data about the catch of a given species in a given water area over a certain time span. In the case of statistics concerning trade, the "trade flow" (import/export) is also included. All statistics collected by FIES are available on the web and accessible by means of an online query panel.


http://www.fao.org/figis/servlet/TabSelector

fisheries fao

Global Capture Production 1950-2006

Contains the volume of fish catches landed by country or territory of capture, by species or a higher taxonomic level, by FAO major fishing areas, and year for all commercial, industrial, recreational and subsistence purpose. more >>

To define your Query, select the items of interest from the selection tabs and choose the entries to show from the Display tab. Use the Hierarchy option to change the grouping of individual items. To Expand/Collapse aggregated items click the plus/minus symbol beside them.

hierarchy: Taxonomic

Path:

- □ ⊞ AMPHIBIA, REPTILIA
- □ ⊞ CRUSTACEA
- □ ⊞ INVERTEBRATA AQUATICA
- □ ⊞ MAMMALIA
- □ ⊞ MOLLUSCA
- ☑ ⊞ PISCES
- □ ⊞ PLANTAE AQUATICAE

HOT TIPS

Click on □ to select/deselect a given node or item.

Click on to select/deselect all children that belong to the given node.

Click on ⊞ or ⊟ to expand or collapse an aggregated node.

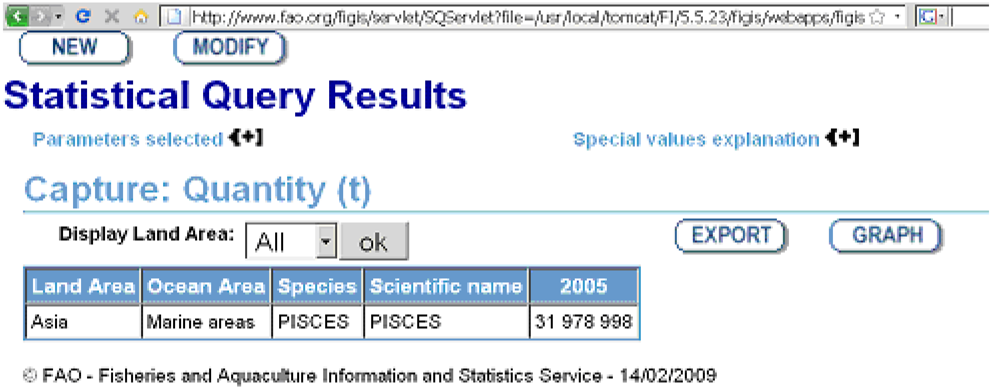
SYMBOLS

= node = some children selected = all children selected

Figure 1. Online query panel for fisheries statistics in FAO.

[8] For explanations about fisheries-related concepts see Annex II.

Figure 1 shows an example of a query that can be composed by using the query panel over the database about “Global Capture Production”. The query composed can be phrased as: “Production of “pisces” in Asia, in the year 2005”. Figure 2 shows the result of that query.



http://www.fao.org/figis/servlet/SQServlet?file=/usr/local/tomcat/FI/5.5.23/figis/webapps/figis

Statistical Query Results

Capture: Quantity (t)

Display Land Area: All

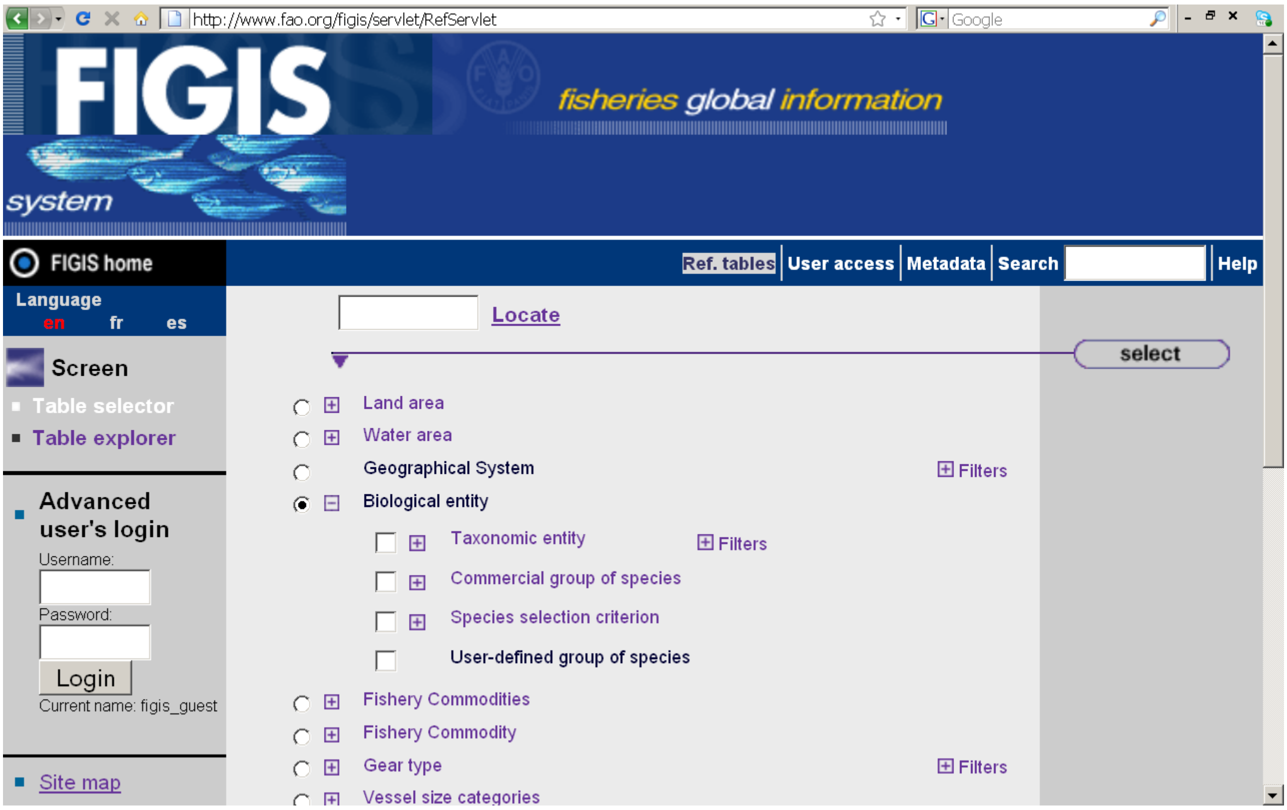
Land Area	Ocean Area	Species	Scientific name	2005
Asia	Marine areas	PISCES	PISCES	31 978 998

© FAO - Fisheries and Aquaculture Information and Statistics Service - 14/02/2009

Figure 2. Result of the query presented in Figure 1.

Data is collected by FIES by means of forms that are sent out on a regular basis, usually yearly, to member countries. The data returned is indexed according to some classification systems that serve as metadata for the statistical pieces of data. The use of standardized classification systems is useful because it is unambiguous (as opposed to using names), and it allows for concise reference to data (within the organization that collects the statistics and in an international level) and easy storage and dissemination. In the domain of fisheries, as in many other domains, many classification systems are available, to account for the fact that (i) different objects (e.g. animals, vessels) are important depending on the environmental area and human culture; (ii) the same objects may be referred to by means of different languages and expressions; (iii) data may be collected differently because of the specific purposes it is going to be used for, and depending on the two mentioned factors. Even in case of international classification systems, variations are often utilized to allow for the specific needs of the institutions using them. It is then common practice to use one classification system of choice for data storing and then convert it into other systems for data dissemination. This explains the many classifications that we are presenting in this deliverable. This should also explain why the reference tables often model coding systems in the domain of fisheries, more than the actual entities in the domain (cf. Sec. 3.1).

The reference tables are also accessible online and represent an important resource of information in the area of fisheries. Figure 3 shows the online searchable visualization of the metadata.



○ ⊞ Land area
○ ⊞ Water area
○ Geographical System ⊞ Filters
● ⊟ Biological entity
- □ ⊞ Taxonomic entity ⊞ Filters
- □ ⊞ Commercial group of species
- □ ⊞ Species selection criterion
- □ User-defined group of species

○ ⊞ Fishery Commodities
○ ⊞ Fishery Commodity
○ ⊞ Gear type ⊞ Filters
○ ⊞ Vessel size categories

Figure 3. Online browser showing the reference tables.

The browser shown in Figure 3 reproduces the internal hierarchical organization of metadata in RTMS. Although for internal use all these pieces of data are usually referred to by means of their ID in the database, for the purpose of data publication and international exchange of data, what really matters is their *code* (as opposed to their names, for example, as natural languages are usually long and they tend to be misunderstood). Reference tables then store codes and their hierarchical structure, when available, and the association between codes and names in one or more languages (usually English, French and Spanish). Correspondence between languages is 1-1 because it results from international agreements (e.g. on names of territories, on commodities classification). More detailed information regarding fisheries statistics can be found in the Handbook of Fishery Statistical Standards [HBFSS] by the Coordinating Working Party on Fishery Statistics (CWP).[9] The entire system that manages the Reference Tables is called Reference Tables Management System (RTMS), of which the core is an Oracle database, called FIGIS.

2.2 Other applications of reference data

The main use of reference data is as metadata for time series. However, they are also used in the context of other applications (just as coding systems have a variety of applications). For example, the fisheries fact sheets [FS] created by FAO[10] do use information about coding systems that are

[9] The Coordinating Working Party on Fishery Statistics (CWP) supported by its participating organizations has served since 1960 as the premier international and inter-organizational forum for agreeing upon common definitions, classifications and standards for the collection of fishery statistics.

[10] See fact sheets about marine resources (i.e. stocks) collected by FIRMS: http://firms.fao.org/firms/inventory/browse; fact sheets about resources, geographical profiles, technologies and information standards available from the FI department web site: http://www.fao.org/fishery/factsheets/en.

stored in RTMS. The fact sheets contents[11] are generated by experts and published electronically as XML documents according to a comprehensive XML schema [FSschema]. In this way, FAO makes available a large amount of information about fisheries, aquaculture and related subjects, including fishing techniques, fishing areas, fisheries and aquaculture country profiles. Fact sheets are grouped by *domains* (e.g. Cultured species, Fishing equipment, Fishery, Gear type), each corresponding to an element under the root FIGISdoc, the root of any fact sheet (XML document). Domains are fully specified by means of nested elements. Each element includes a description meant for human use.

The schema makes use of existing standard element sets such as Dublin Core [DC], Extended Dublin Core [EDC], AGMES [AGMES] and AIDA [AIDA]. It also incorporates, wherever possible, existing classification schemes (such as ISO standards for countries, currencies, languages, and other fisheries-related international classification schemes) most of which are stored in the RT.

It is important to note that the schema was conceived as a means for editors to create structured documentation, and as such was not created based on a relational or ontological model, but was rather organized following hierarchical document formatting conventions. A dictionary of the elements used in the schema is available online [FSdic].

2.3 Biological entities

Aquatic species of interest to FAO (both for fisheries and aquaculture) are organized and maintained in the Aquatic Science and Fisheries Information System (ASFIS)[12]. Species are provided with a number of codes: taxonomic codes, ISSCAAP and 3-alpha. Taxonomic codes reflect a biological point of view (simplified taxonomic classification). The ISSCAAP system reflects a commercial point of view, while 3-alpha is a coding system for data exchange, developed by UN agencies. An English name is available for most of the records, and about one third of them also have a French and Spanish name.[13] Currently, the ASFIS list includes nearly 11,000 species items selected according to their interest or relation to fisheries and aquaculture.

2.3.1 Taxonomical classification of species in FAO statistics

Since Linneus the taxonomic classification of organisms is widely used and it has become perhaps the best known example of classification system. Although the original classification of Linneus has largely changed, the idea of classifying species[14] into a ranked hierarchy is still widely used. Currently, the scientific taxonomy of species usually includes the following ranks:[15]

Life – Domain – Kingdom – Phylum – Class – Order – Family – Genus – Species.

However, not all taxa are used for all species and for all purposes. In particular, for management purposes a simplified taxonomy may be used. The simplified taxonomy used in FAO consists of:

[11] The actual fact sheets are usually the result of a complex procedure to make the content provided by the experts available in a format suitable for publication, after integration of information stored in other systems (e.g. geographical maps).

[12] http://www.fao.org/fishery/collection/asfis/1

[13] Member agencies of the CWP have agreed to use these standard species names in statistical publications and questionnaires. However, (a) it has not been possible to assign appropriate names in all three languages to all species items, and (b) these names may not correspond with nationally or regionally-used common names.

[14] Also the concept of species has changed over time. For a long time species have been defined on the basis of morphological similarities, then the idea of compatibility in breading has been used to distinguish species, currently species are defined as a separately evolving lineage that forms a single gene pool.

[15]The taxonomy may be detailed so as to include many more ranks, such as Subkingdom, Subphylum, Subclass, Suborder and so on. Also, note that there are slightly different ranks for zoology and botany.

Main group – Order – Family – Genus – Species.[16]

The taxonomic classification is encoded in a taxonomic code that is often also used as an identifier. Starting from a species, the taxa above are added in the ASFIS list. However, since only “useful” taxa are added, some taxonomic chains may be “incomplete”.

A taxonomic code for biological entities is a 10-digit code (but in a few special cases three extra digits have been added) that for any entity specifies its *type,* i.e. if it is a major group, an order, a family, a genus or a species, and its complete hierarchical path. Table 1 shows an example of taxonomic code (of the species “Swordfish”: 1750400301) to show how the 10 digits of the taxonomic code are organized.

	Main grouping	*Order or* *high taxonomic level*	*Family*	*Genus*	*Species*
Digits	digit 1	digits 2 and 3	digits 4 and 5	digits 6, 7, 8	digits 9, 10
Example	1	75	04	003	01

Table 1. Structure of the 10-digit taxonomic code used for biological entities.

Note that this is an example of an incomplete taxonomy, since the family of Swordfish (which should have taxonomic code 17504XXXXX) is not present in the list, while the corresponding order is “Tuna-like fish nei”, or “Scombroidei” (code 175XXXXXXX).

2.3.2 ISCCAAP grouping of species

The ISSCAAP code is assigned according to the FAO 'International Standard Statistical Classification for Aquatic Animals and Plants' (ISSCAAP) which divides commercial species into 50 groups on the basis of their taxonomic, ecological and economic characteristics.[17] ISSCAAP classification system is based on a group of species (ISSCAAP group) that may be further grouped into divisions (ISSCAAP divisions). The ISSCAAP codes have been assigned to all the species items included in the ASFIS list except the sea birds and the sea snakes as these animals are not included in any ISSCAAP group. ISSCAAP groups may only contain species; but some species may not be part of any ISSCAAP group. No species belongs to more than one ISSCAAP group. ISSCAAP divisions only contain ISSCAAP groups.

2.3.3 Inter-agency 3-alpha code

The "Inter-Agency 3-alpha Code" was developed by the CWP to establish a common system to exchange data among its members and facilitate the reporting of fishery statistics from national correspondents. This coding does not presuppose any grouping nor does it assume any taxonomic classification. Once a 3-alpha identifier has been assigned to a species item it is not changed and thus it is a permanent reference to a species item. Codes for species items which have been cancelled (often because the item is recognized as a synonym of a valid species) are not reused but are considered as 'dead codes'.

[16] Genus is in the ASFIS list but it is not used in RTMS.

[17] An interactive browser of the correspondences between ISSCAAP groups and ASFIS species is available: http://www.fishbase.org/report/ISSCAAP/ISSCAAPSearchMenu.cfm

In the first years of the 3-alpha code system, the three letters chosen were related either to the scientific or the English name of the species items. For items taxonomically classified as above the species level, an 'X' was typically used as the third letter of the combination. With the increased number of codes assigned, these criteria became mostly inapplicable.[18]

2.3.4 Summary

- The list of species relevant to fisheries is originally maintained in the ASFIS list.
- All species are provided with a taxonomic code based on their taxonomic classification using 5 taxa: group, order, family, genus, species.
- The taxon genus is not used as reference for time series.
- Taxonomic chains may be incomplete (e.g. there may be groups with no specification of order or family, and species with none of the taxonomic specification mentioned above).
- ISSCAAP is a classification of species according to their commercial value, therefore it is also considered a classification of fisheries commodities.
- ISSCAAP groups only contain species.
- There are species in the ASFIS list that do not belong to any group.
- ISSCAAP divisions only contain ISSCAAP groups.
- 3-alpha codes are used for quick identification and data interchange; it does not have any correspondence with biological classification.
- 3-alpha codes are given to all biological entities classified with a taxonomic code, but also to some ISSCAAP groups.

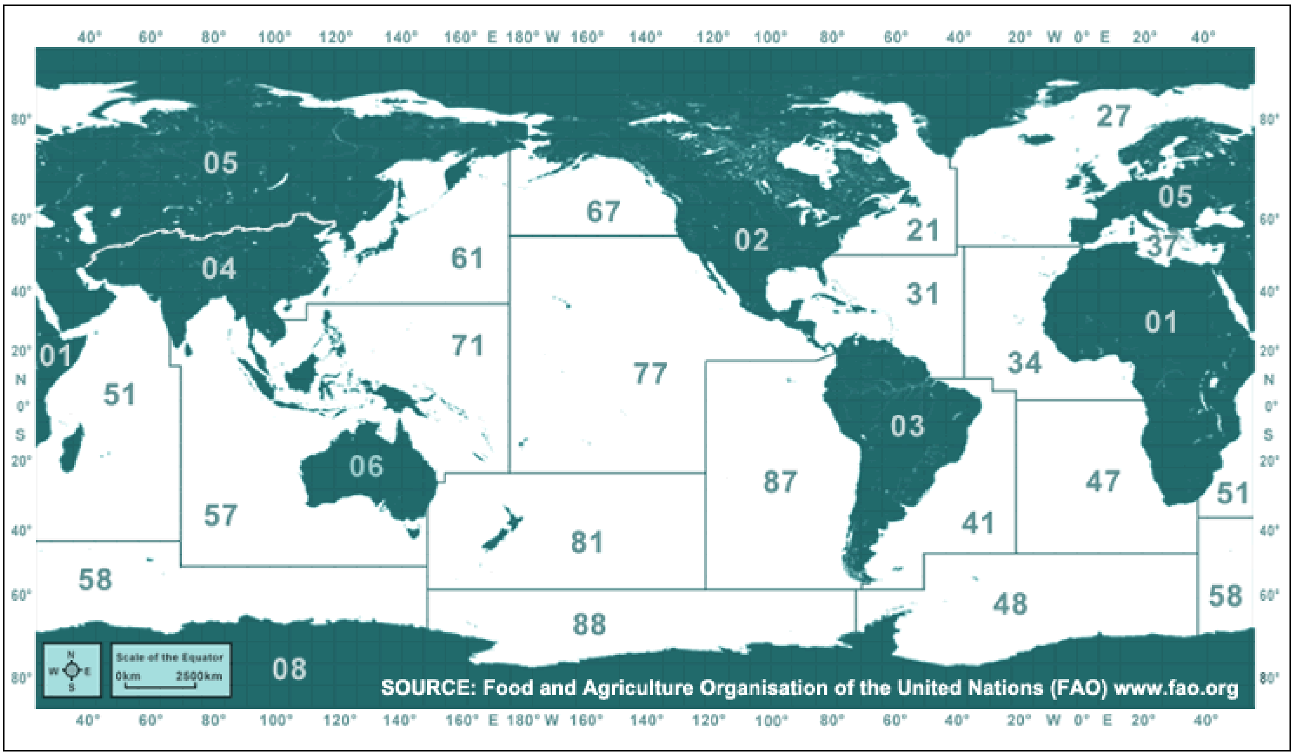
2.4 Water areas

Marine and inland waters are divided into a variety of zones and areas, depending on the purpose of the division (e.g. legal jurisdiction, statistical reporting of catch data, environmental assessment) and on the author of the division (national or international body).

2.4.1 FAO divisions

For the purpose of data collection and statistical reporting, the water areas in the world are organized into a system of 27 FAO areas also called “FAO division areas”, which consist of major areas, divided into sub-areas, each divided into divisions, and these finally into sub-divisions. Figure 4 shows the FAO areas only. As one can see from Figure 4, the FAO divisions cover marine waters (e.g. major areas number 61, 88, and 37) as well as inland waters (e.g. major areas number 06, 02, 08). For the entire list of FAO divisions, see [FAOdiv].

[18] The CWP includes new identifiers wherever an authority, national, international or other, considers them to be of use. However, the CWP does request that all potential users consult FAO, and in any case, never use an identifier that is not in the current list without the prior approval of FAO (CWP, 1990).


40° 60° 80° 100° 120° 140° 160° E 180° W 160° 140° 120° 100° 80° 60° 40° 20° W 0° E 20° 40°
80° 60° 40° 20° N 0° S 20° 40° 60° 80°
05
04
01
51
57
58
08
61
71
06
67
77
81
88
02
21
31
03
87
41
27
05
37
01
34
47
51
48
58
N W E S
Scale of the Equator 0km 2500km

SOURCE: Food and Agriculture Organisation of the United Nations (FAO) www.fao.org

Figure 4. FAO division of water areas.

The FAO division of water areas forms a strict and complete hierarchy based on inclusion, or part-of. Water areas have names in natural language only at the area level, while internal divisions are given numeric names.

The codes assigned to each division are taxonomic[19] so as to keep track of the relation of inclusion between areas and sub-areas. For example, the major area “SouthEast Pacific” has code 87, and one of its subdivisions has code: 87.2.1.1. Codes do not contain any indication of whether a water area is inland or marine (although inland waters are given a continuous numeration from 01 to 08).

2.4.2 Large marine ecosystems (LME)

Large marine ecosystems (LMEs) are regions of the world's oceans, encompassing coastal areas from river basins and estuaries to the seaward boundaries of continental shelves and the outer margins of the major ocean current systems. They are relatively large regions on the order of 200,000 km^2 or greater, characterized by distinct bathymetry, hydrography, productivity, and trophically dependent populations.

The system of LMEs has been developed by the US National Oceanic and Atmospheric Administration (NOAA)[20] to identify areas of the oceans for conservation purposes. According to this system, 64 LMEs have been identified (see Figure 5). For 33 of them, NOAA has conducted studies of principal driving forces affecting changes in biomass.

[19] Any library user is familiar with taxonomic codes because of the Dewey classification system. In that system the digits composing the number do not signify numbers, but should be interpreted according to the classification system.

[20] http://www.lme.noaa.gov/

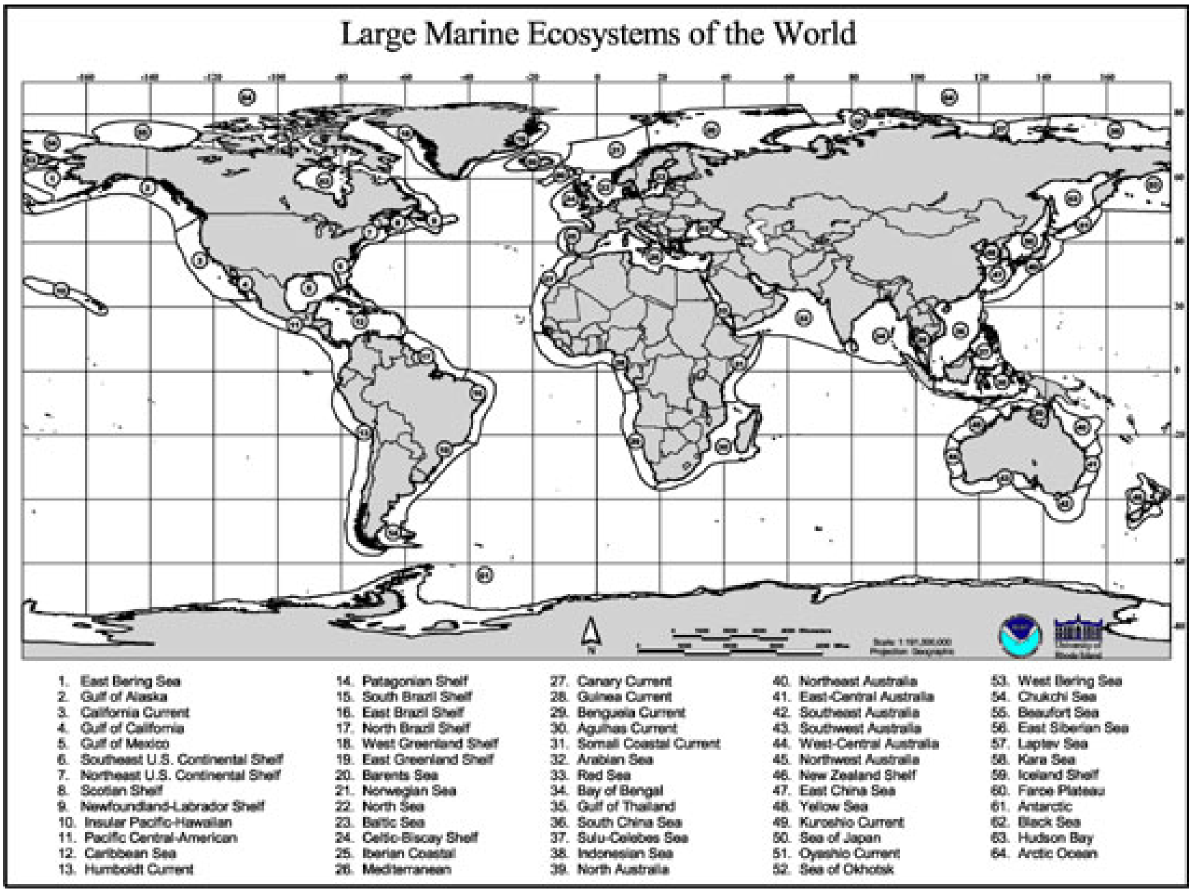


Figure 5. Map of the large marine ecosystems identified by NOAA.

Although the LMEs cover only the continental margins and not the deep oceans and oceanic islands, the 64 LMEs produce 95% of the world's annual marine fishery biomass yields. Most of the global ocean pollution, overexploitation, and coastal habitat alteration occur within their waters.

An example of a large marine ecosystem is the Humboldt Current, a cold, low-salinity ocean current that flows north-westward along the west coast of South America from the southern tip of Chile to northern Peru, in the direction of the Equator (listed as number 13 in Figure 5). The Humboldt Current is one of the major upwelling systems of the world, considered the most productive marine ecosystem in the world. It supports an extraordinary abundance of marine life and the world’s largest fishery.

2.4.3 Exclusive Economic Zone (EEZ)

Under the law of the sea (Third United Nations Convention on the Law of the Sea, concluded in 1982), an Exclusive Economic Zone (EEZ) is a sea zone over which a state has special rights with regards to the exploration and use of marine resources.

Generally, a country's EEZ extends to a distance of 200 nautical miles (370 km) out from its coastal baseline, perpendicular to the baseline. However, when an overlap occurs (i.e. when state coastal baselines are less than 400 nautical miles apart), the actual boundary is delineated by the countries (generally, any point within an overlapping area defaults to the most proximate state). Figure 6 depicts a typical sequence of layers of water divisions from the point of view of legal jurisdiction starting from a country’s coast baseline (i.e. territorial waters, contiguous zone, exclusive economic zone, continental shelf, international waters).

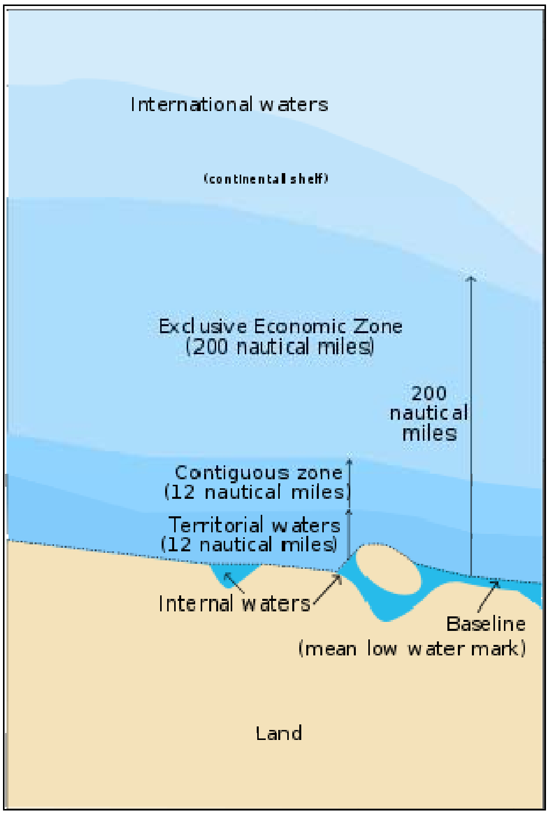


Figure 6. Water division from the point of view of country's legal jurisdiction.

Thus, the EEZ overlaps both the contiguous zone and territorial waters. The importance of EEZ cannot be overestimated, as can be witnessed by the disputes between states over marine waters.

Fisheries management, usually adhering to guidelines set by the FAO, provides significant practical mechanisms for the control of EEZs. Trans-boundary fish stocks are an important concept in this control. Trans-boundary stocks are fish stocks that range in the EEZs of at least two countries. They can be contrasted with straddling stocks, which range both within an EEZ as well as in the high seas, outside any EEZ. A stock can be both trans-boundary and straddling.

From the point of view of data collection and fisheries management, it is important to map the overlap between EEZ and FAO divisions used for statistical data reporting.

2.4.4 Summary

- FAO divisions form a part-of hierarchy with 4 levels: areas, subareas, divisions, subdivisions.
- The hierarchy of FAO divisions is always complete (i.e. an area may not have subareas, but all subareas have an area to which they belong).
- Each FAO division is given a taxonomic code.
- The distinction between inland and marine water is not included in the FAO divisions.
- FAO areas overlap with other types of divisions of water areas, such as large marine ecosystems and exclusive economic zones.

2.5 Land areas

Land territories are central to most statistical collections. Fisheries data is no exception as in 1954 the United Nations Statistical Commission decided that fish catches should be assigned to the country of the flag flown by the fishing vessel.[21] It should be noted that "flag country" generally refers to the country in which the vessel (which may be small and not physically flying a flag) is registered.

The names of territories (countries and groups) are established by international agreements. By agreement, two types of names of territory are given in each language: long names to be used in official documents, and short names to be used in informal communications.

Codes commonly used for land areas include those maintained by the International Standard Organization (ISO) (ISO-3166 ALPHA-2 [ISO2] and ALPHA-3 [ISO3]), by the United Nations and its agencies such as the United Nations Development Programme (UNDP) and the UN Statistical Division, and by FAO (FAOSTAT[22], GAUL and FAOTERM). The reason for having so many coding systems is that each of them tends to be specific to the purpose and application for which they are used. For example, some coding systems also cover land areas below the national level, others above it.

Data can then be aggregated above the national level into groups defined according to different criteria, such as a geographic or economic unit. Continents, such as Africa and Asia, are typical geographical regions; the Caribbean Community (CARICOM), the Union Economique et Monetaire Ouest Africaine (UEMOA), and the Gulf Cooperation Council (GCC) are examples of economic regions.[23]

Since territories and groups change over time, the RTMS also include their range of “validity”, in order to continue to be able to query the statistical database according to territories no longer existing. This situation occurs when territories join together, as in the case of East Germany and West Germany (both created in 1949 and dissolved in 1990 to become Germany), or split, as in the case of Serbia and Montenegro (that in 2006 split into Serbia *and* Montenegro). Groups of territories are also dynamic, as geographical groups (continents) “change” when their member territories change, and economic groups similarly change every time a country joins or leaves a group.

2.5.1 Summary

- Countries names are established by international agreements
- A variety of coding systems is used to identify them.
- Countries may be member of groups (e.g. geographical, political, economic)
- Countries do change over time: they can merge or split.

[21] This is not valid anymore, but old data is collected in this way.

[22] http://faostat.fao.org/

[23] For a list of regional economic organizations with which FAO works, the reader can refer to [FAO-groups].

2.6 Fish stocks[24], aka aquatic resources

From a biological point of view, a stock comprises all the individuals of fish in an area, which are part of the same reproductive process. A stock is self-contained, with no emigration or immigration of individuals from or to the stock. It occupies a well-defined spatial range and is independent of other stocks of the same species. Random dispersal and directed migrations due to seasonal or reproductive activity can occur. Some species form a single stock (e.g. southern bluefin tuna) while others are composed of several stocks (e.g. albacore tuna in the Pacific Ocean comprises separate Northern and Southern stocks). The impact of fishing on a species cannot be determined without knowledge of this stock structure.

On the other hand, and from a pragmatic point of view, a ***stock*** is the part of a fish population which is under consideration from the point of view of actual or potential utilization. The term *"resources"* is often used when referring to vaguely defined "stocks", especially for management purposes. The FAO Glossary for Responsible Fisheries indicates that ***aquatic resources*** are "biotic elements of the aquatic ecosystem, including genetic resources, organisms or parts thereof, populations, etc. with actual or potential use or value for humanity. Fishery resources are therefore those aquatic resources of value to fisheries". Just like a stock, a fishery resource is defined in space and its geographical demarcation and often has a political or jurisdictional connotation (e.g. Moroccan resources; Exclusive Economic Zones (EEZ) or high seas resources). Potential or actual catch is a typical resources indicator reflecting the notions of use or value attached to the resource's concept. It could be local, national (e.g. the shrimp resource of Ghana), regional (e.g. Atlantic tuna) or global (cephalopod resources of the world).

Of course, collecting catch data on a given stock is easiest when it is harvested by only one country, but this rarely happens because stocks often cross political boundaries and many countries are generally involved in their exploitation. For example, many fish follow migratory routes that take them through the waters of several countries or out into international waters. The situation may be further complicated by the life history of the target species. In many marine species, the eggs and larvae drift away from the spawning grounds into the open sea. In others, young and adult fish spend most of their life in the open ocean, returning to coastal waters or even rivers only to breed. So the success of a fishery or the outcome of a management scheme in one area can be affected by catches of another some distance away. For example, attitudes to an expanding salmon fishery off Greenland changed rapidly when it was discovered that the fish being caught there came from, and could have been expected to return to, American and European rivers.

Differences between stocks may be established by taxonomic, physiological and biochemical features (e.g. the analysis of blood proteins has been used to differentiate between the many cod stocks that exist in the north Atlantic). Tagging and fin clipping are also widely used to track the migration and distribution of fish. From a practical standpoint, deciding on the practical boundaries of a stock and of a management scheme is less critical when the pattern of fishing is relatively uniform over a wide area. In such a situation, the fish of a given species may be treated as a single unit stock if they do not differ greatly in reproduction, growth and mortality rates. On the other hand, if the intensity of fishing differs markedly within the area occupied by the species and the fish populations in different parts of the area rarely mix, then the fish population in each part would have to be treated as a unit stock.

While there is a substantial amount of information on important stocks exploited by large scale industrial fishing, there is little or no information on a very large number of smaller coastal stocks

[24] The information presented in this section is based on the following sources: Fisheries Resources Monitoring System (FIRMS) http://firms.fao.org/firms; UN Atlas of the Ocean http://www.oceansatlas.org/index.jsp; ASFA thesaurus http://www4.fao.org/asfa/asfa.htm.

(e.g. exploited by small-scale fisheries) or very deep sea stocks exploited without control far away from the coasts (e.g. on sea-mounts). In particular, the total number of entities identified as stocks in the world is not known and no comprehensive classification exists. FAO has started the development of a worldwide inventory of "stocks" in the sense of distinguishable, pragmatic resource units identified by governments as management units and often, as a consequence, used as a basis for stock assessment.

2.6.1 Summary

- A stock may be multi-species.
- Some species form a single stock while others are composed of several stocks.
- A stock can be highly migratory, straddling or shared. This affects the type and number of water areas in which they can be caught.
- A stock may be global, ocean-wide, regional, national and local, depending on the distributions on the planet.

2.7 Fisheries commodities

A commodity is anything for which there is a demand and a market,[25] and which is supplied without qualitative differentiation across the market – i.e. primary goods such as oil, gold, sugar, milk, copper, rice, fish.

Fish is as a highly perishable commodity, and often undergoes treatments which prolong its shelf life and quality as food. Fish is also a very widely traded commodity. When considering statistical aspects related to fish and fish products in the fishery industry as a whole, one is faced with a wide variety of raw fishery materials, semi-processed and fully-processed commodities, crossing all the various fishery phases. The physical magnitude and value of the intake and output of the different kinds of fishery commodities can be measured in specified periods of time – such as days, weeks, seasons, or years. Statistics covering any of the above phases must be dovetailed, linked or integrated and the first indispensable step is an adequate fishery commodity classification. The classification can be used as statistical standard for more than one statistical system, e.g. the trade system, industrial censuses, censuses of commercial and service establishments, wholesale and retail price systems. Several coding and classification systems are available for fisheries commodities.

The **International Standard Statistical Classification of Fishery Commodities (ISSCFC)** [ISSCFC] is used for detailed information on countries or zones. The ISSCFC is a taxonomic classification system maintained by FAO and used to collect data on commodities from countries. Its maximum depth is six levels. It is an expansion of the United Nations Standard International Trade Classification, revision 3 (SITC, see below). The ISSCFC classification is linked to the Harmonized System of the World Customs Organization (see below), and to the ISSCAAP classification (see Sec. 2.3.2).

The **Standard International Trade Classification (SITC)** coding system is developed by the United Nations' Statistical Office on the basis of earlier international work on the subject. Its first version, in 1950, was based on the League of Nations' Minimum List of Commodities for International Trade Statistics, which was published in 1937. Since then it has been revised twice and the current version is revision 3 [SITC3]. SITC reflects various aspects of commodities

[25] However, not everything that can be sold is part of the commodity trade (no future market exists for them), such as asphalt, fresh and cut flowers, tomatoes, figues, etc. Generally for these products the only price information available is based on information from producers, consumers and traders.

including the materials used in production, the processing stage and the importance of the commodities in terms of world trade. It has a hierarchical structure consisting of Sections, Divisions, Groups, Subgroups and Items. The SITC coding system is available in the following languages: Arabic, Chinese, English, French, Russian, and Spanish. Only the necessary fragment of SITC is used in FAO for fisheries commodities.

The **Harmonized System (HS)** [HS07] was introduced in 1988 by the World Customs Organizations (WCO)[26], and has become an internationally accepted method of classification for traded products. The HS is “harmonized” in relation to the classification of the UN and the European Community. Goods are classified according to simple objective criteria and applications. The HS, a revision of the CCCN (Customs Cooperation Council Nomenclature) classification system, includes a six-digit sub-heading that was introduced for more precise tagging of products.[27] It is intended to serve as a universally accepted classification system for goods so countries can administer customs programs and collect trade data on exports and imports. It was designed to replace the varied tracking methods used by countries and create one common classification system with which to track trade and apply tariffs. The basic system is a 3-level taxonomic code forming a 6-digit number identifying basic commodities. Each country is allowed to add additional digits for statistical purposes (called HS-4). For fisheries commodities, FAO uses a fragment of HS-4. In the Harmonized System articles are grouped largely according to the nature of the materials of which they are made, as has been traditional in customs nomenclatures. The HS contains approximately 5000 headings and subheadings covering all articles in trade.

2.7.1 Summary

- The basic classification for FAO is ISSCFC.
- ISSCFC is linked to HS and ISSCAAP.
- ISSCAAP may be looked at as both a classification of species and a classification of commodities.
- ISSCAAP is then linked both to ISSCFC and the taxonomic classification for species.

2.8 Fishery fleet

The term "fishery fleet" or "fishery vessels" refers to mobile floating objects of any kind and size, operating in freshwater, brackish and marine waters which are used for catching, harvesting, searching, transporting, landing, preserving and/or processing fish, shellfish and other aquatic organisms, residues and plants.

The term "fishing vessel" is used instead when the vessel is engaged only in catching operations. "Non-fishing vessels" are those vessels performing other functions related to fisheries, such as supplying, protecting, rendering assistance or conducting research or training.

In order to assess fleet capacity, it is necessary as a bare minimum to have estimates of vessel numbers and main vessel characteristics. If the fleet consists of only one type of vessel, the number of fishing vessels can be used to express the total fishing power or capacity of the fishing

[26] The system was originally developed by the Customs Cooperation Council (CCC), now known as the World Customs Organization (WCO). The WCO, located in Brussels, is an international organization consisting of representatives from about 139 countries and territories.

[27] At present this system contains 21 sections, 97 chapters and 1,241 headings at the four-digit level, 930 of which are further divided in sub headings. HS-1996 (revision 1) represented a total of 5,113 separate categories of goods identified by a six-digit code. Most of the countries that have adopted HS have added one or more digits to further classify products of particular national interest (8-digit or 10-digit level).

fleet. If the fleet consists of vessels of different designs, any survey to determine the capacity of a given fishing vessel would need to collect information on a number of vessel characteristics. Gross tonnage (GT), length and engine power would be amongst the most important characteristics, and it is likely that gross tonnage would be the most important single variable influencing fishing capacity (see Sec. 2.9 below).

2.9 Vessel types and size

In order to assess fleet capacity it is necessary as a bare minimum to have estimates of vessel numbers and main vessel characteristics, such as the vessel type and its size or length.

In international law, as well as in practice, several systems of tonnage measurement have existed side by side. Traditionally, records of measurements of a ship's size were expressed in tons of 100 cubic feet each called Gross Register Tonnage (GRT), as defined by the Oslo Convention (1947). Tonnage was used as a basis for taxes, berthing, docking, and passage through canals and other facilities. However, the method of tonnage measurement has evolved and differs considerably from country to country. A number of international meetings on the subject concluded with the International Convention on Tonnage Measurement of Ships (London, 1969). The Convention, commonly known as the 1969 Tonnage Convention, entered into force in July 1982, though existing ships were not required to comply with the Convention until July 1994. At that time, Gross Tonnage (GT) as defined by the 1969 London Convention became obligatory for all vessels of 24 metres in length and over which were engaged in international voyages.

Although the London Convention has been adopted for vessels of 24 meters in length and over, for many vessels only data conforming to the Oslo Convention are available. The situation varies from country to country. The two conventions produce very different tonnage values: although GT measurement is higher than GRT, there is no simple correlation between the two units (GT is often double the GRT, but sometimes as much as four times the GRT).

Based on the international convention in use, FAO fleet data on the vessel tonnage is measured according to the Oslo Convention (1947) expressing data by GRT [ISSCFVgrt] until 1995; and according to the London Convention (1969) expressing data in GT since 1996 [GT].

As for the type of vessels, the International Standard Statistical Classification of Fishery Vessels by Vessel Types (ISSCFV), based on the type of gear used by the vessels, approved by the CWP in 1984 is adopted [ISSCFVgrt].

Starting with the collection of data for 1996, several other changes were implemented in the form used to gather data: non-fishing vessels were excluded from the inquiry, numbers and capacity data are now collected for broad groups of fishing vessel types, and length has been defined as the main characteristic of measurement in international data collation. Discussions are ongoing within the CWP on the possibility of further improvements to the ISSCFV classification "by type" to reflect the state of current technology developments.

2.9.1 Summary

- Vessels are usually classified according to two dimensions: size (expressed by tonnage or class power, i.e. length) and type.
- Main classifications by tonnage are: GT and GRT.
- Vessels need not be classified according to both dimensions.
- Classifications are not exclusive.
- ISSCFC is both about size and type of vessels.

2.10 Gear types

The type of gear installed on a vessel determines the type of fish that can be caught, therefore it is often used in statistical collections to determine the fleet power. The main classification of gear types is the International Standard Statistical Classification of Fishing Gear (ISSCFG), adopted in 1980 during the 10th Session of the CWP [ISSCFG].

Although this classification was initially designed to improve the compilation of harmonized catch and effort data questionnaires and in fish stock assessment exercises, it has also been found to be very useful for fisheries technology and the training of fishermen. It has been used in particular for reference in works dealing with the theory and construction of gear and for the preparation of specialized catalogues on artisanal and industrial fishing methods. The classification of gear is used in FAO only for the compilation of fishery fact sheets.

3 The first network of fisheries ontologies

In this chapter, we present in detail the data on which the network of ontologies is based. In Sec. 3.1 we describe the database of reference tables. In Sec. 3.2 we describe the process and the outcome of the validation and revision of the ontologies produced in D7.2.2. In Sec. 3.3 we deal we the reengineering of the ASFA thesaurus, and present sample mappings between ASFA and the ontologies based on RTMS. In Sec. 3.4 we describe the process followed to create the network of ontologies, paying special attention to the extraction and modelling of the data used to link the ontologies. In Sec. 3.5 we present the first network of fisheries ontologies, while in Sec. 3.6 we highlight its salient features.

3.1 The database of reference tables

All entities relevant to the fisheries domain are organized into *types* and characterized by *meta codes* (e.g. *species* is a type of fisheries entity, with meta code = 31005). Meta codes are hierarchically organized, according to an ‘is a’ (or ‘extend’) type of relationship: for example, a Species type is a (extends) Taxonomic entity (meta= 31000), which in turn is a (extends) Biological entity (meta=30000). The hierarchy of types are strict (no multiple parenthood allowed); it is represented in a dedicated table. Table 2 presents the highest levels of the hierarchy of types. The hierarchy is presented in detail in Annex V.

- 1 figis object
    - 10 000 Land area
    - 20 000 Water area
        - 21 000 Environmental area
        - 22 000 Fishing Statistical area
        - 24 020 Jurisdiction area
    - 30 000 Biological entity
        - 31 000 Taxonomic entity
        - 32 000 Commercial group of species
    - 45 000 Fishery commodity
    - 50 000 Gear type
    - 60 000 Vessel size categories
    - 64 000 Vessel type

Table 2. A fragment of the hierarchy of types representing the first 3 levels.

Any instance of any type (a given species, a territory, a gear type, an ISSCAAP group) is an *item* represented as a row in a table and characterized by its meta. Although everything can be an item, and the ultimate objects of the reference data are real world objects, the substance of the reference system is *classification systems*. For example, a country (defined and identified according to political agreements), a continent, species and families of species (as the two taxa in a taxonomic classification of living organisms), and names of ISSCAAP groups (e.g. “Tuna, bonitos, billfish" with ISSCAAP code 36) are all items in the database. Information about individual items is usually organized into one table per domain.

The fact that species are classified taxonomically into families (and then into orders and groups) is rendered by means of the group-member mechanism, so that a family is a group the members of which are species.[28]

The group-member mechanism is also used to describe relations between entities the type of which belongs to separate branches of the hierarchy of types. For example, the classification of species within the ISSCAAP classification is modelled as an alternative way of classifying biological entities (they are siblings in the hierarchy of metas, Annex V) with respect to the taxonomic classification. Hence no relationship is stored in the hierarchy of meta. However, there is a relation between ISSCAAP and species, being that every ISSCAAP group contains at least one species (and only species); while some species may not be included in any ISSCAAP group. This relationship is modelled as a group-member relation, stored in a dedicated table.

The mechanism just described is very flexible as it allows one to add any type of relations between entities, without touching the classificatory scaffold of the database. The drawback of this is that things soon get complicated, and without an accurate documentation (often difficult to get when data is manipulated within a tight schedule) it is difficult to keep track of the exact information stored in the DB.

The entire hierarchy of meta codes is stored in one *meta table* (called md_refobject in the database). Data concerning each domain is then organized into two tables: [29]

1. one *item table*, where all items in the domain are listed, together with all pieces of information attached to them (e.g. names, codes, meta), and
2. one *group table*[30], in which the actual hierarchy is stored by means of the group-member structure. It may also contain groupings of entities that belong to different branches of the hierarchy of types. This table uses foreign keys to the group and member involved, from the item table, together with the meta code of the group.

Since the whole list (hierarchy) of meta codes is stored in a dedicated table, all item tables and group tables refer to it by means of foreign keys.

All hierarchies within a domain can then be unrolled by looking at a total of three tables: the meta table, plus the item table and the group table(s) corresponding at the domain at hand. For example, in order to get and interpret all reference data concerning biological species one needs to look at the meta table md_refobject, at the item table called fic_item, and at the group table called fic_item_grp (Figure 7).

[28] In fact, given that taxonomies of species may be "incomplete", the group – member mechanism extends to describe *all* groups: (group, order) (group, family) (group, species) (order, family) (order, species) (family, species).

[29] Note that there is no table in the database called *meta table*, *item table* or *group table.* This terminology is only used to help the reader grasp the high level structure of the database.

[30] This may be either a table or a view.

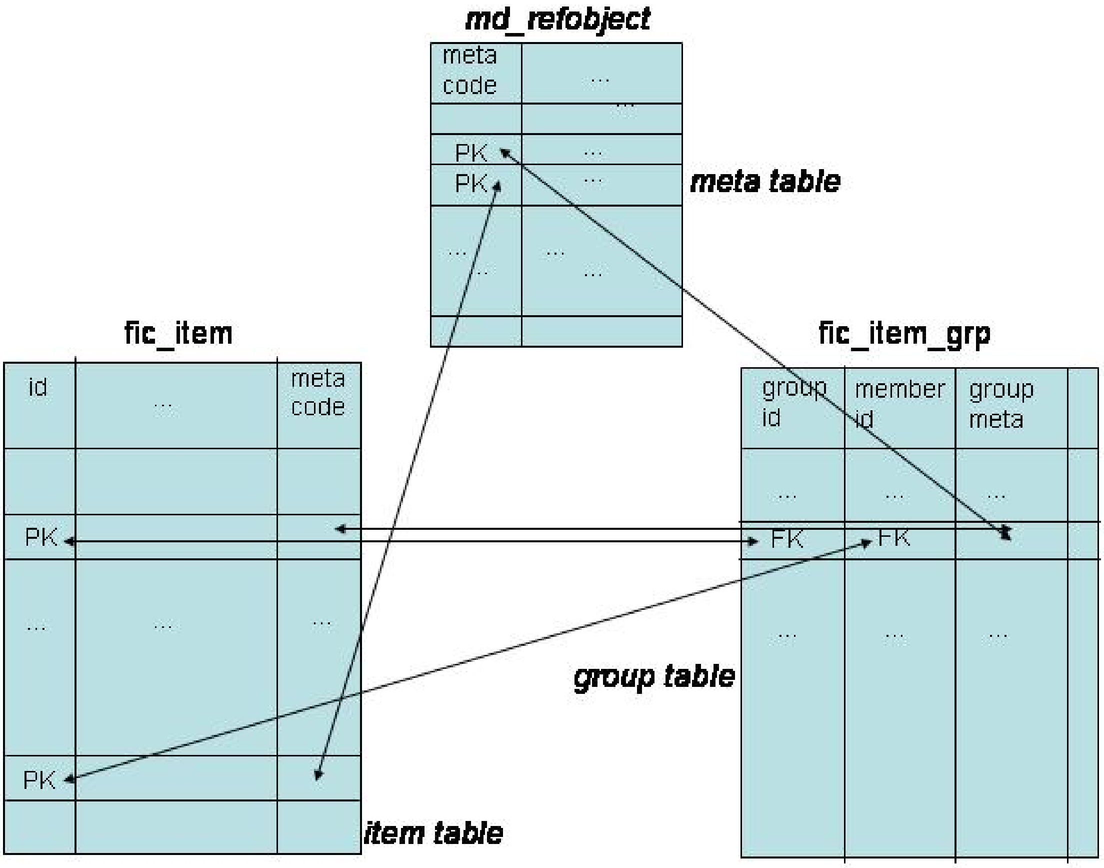


Figure 7. The FIGIS database: tables for the domain of biological entities.

In some cases, special types of meta codes, called *filters*, are used. The only difference between a meta code and a filter is the following. A meta code is associated with each item in the database, therefore meta codes appear in the three tables mentioned above. A filter is a meta code that is not associated with any item. Filters are only used to create hierarchies, and therefore they appear in the meta table and in the group tables, but not in the item tables. Filters may be recognized by a flag[31] in the table.

Since the group table only contains pairs of codes corresponding to group-member association, in order to reconstruct any hierarchy deeper than two levels it is necessary to apply self joins (Figure 8: note that elements may appear both as group and member (e.g. ID of M1= ID of G2).

[31] ismajor=2

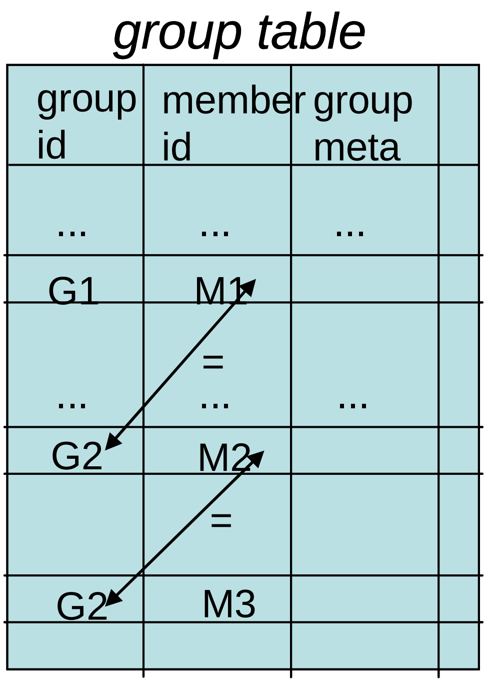


Figure 8. Typical structure of a group table.

Moreover, in order to be able to associate the information stored in the item table with the hierarchical information stored in the group table, it is necessary to apply left and inner joins.

3.2 Revision of the ontologies released in D7.2.2

Some of the ontologies included in the network are new versions of the ontologies presented in D7.2.2, resulting from a process of validation and revision that we describe here.

Based on the RTMS, we released six ontologies as part of Deliverable D7.2.2, all of which are still available online.[32] A detailed description of their model was given in D7.2.2, Sec. 7. After the release of the ontologies in D7.2.2, a cycle of validation and revision started. We received comments from fisheries experts, CNR and during the EC project review. Comments from fisheries experts were gathered by means of a questionnaire developed by FIES and the KCEW group, followed by a face-to-face meeting in order to clarify and expand upon results. The questionnaire was about the user role/work activities, completeness and correctness of the ontologies. Ontologies were reviewed by a domain specialist, a fisheries software developer, and a fisheries information manager using Protégé[33] 4.0 Alpha and OWLDoc[34] for Protégé.

The comments gathered from the three sources just mentioned (FAO/FIES, CNR, EC review) can be summarized in the following way.

3.2.1 Content of the ontologies and appropriateness to domain

Fisheries experts found that the ontologies correctly reflect the domains at hand. Also the level of modularity was found appropriate, although they found that it could be taken even further, for

[32] http://www.fao.org/aims/neon.jsp

[33] http://protege.stanford.edu/

[34] http://protege.stanford.edu/doc/owl/getting-started.html

example by distinguishing even more carefully between coding systems and the correspondences between them. We accommodated these comments especially for that which concerns the water areas and commodities.

For the second revision of the ontologies, they suggested that some branches of RTMS should be further developed by adding more information, especially by adding ISSCAAP classifications for commercial species of fish, and including geopolitical information into the ontology of land areas. These suggestions were accommodated by reengineering the ISSCAAP classification and also linking it to both the taxonomic classification of species and the ISSCFC classification of fisheries commodities. Moreover, in agreement with FIES, we have started experimenting with the inclusion into the network of an ontology developed in FAO but independently of NeOn: the “geopolitical ontology”[35] [KIM09] which includes detailed geopolitical information about land areas (countries and group of countries). The geopolitical ontology is a fundamental component of the “FAO country profiles”[36].

Experts also provided indication of other fisheries domains to render as ontologies, in particular stocks, and *fisheries*. The current version of the network does include an ontology on fish stocks, while the concept of fisheries will be included in the next revision of the network.

By using NTK and the OWLDoc[37] plug-in to show the ontologies to fisheries experts we also gained useful feedback about usability and functionalities of the software, which contributed to making improvements to the plug-in. One interesting comment was that related classifications were not accessible from OWLDoc.

3.2.2 General modeling issues

We received several questions concerning the focus of the ontologies. In fact, on the one hand, the names given to ontologies and classes suggested that the ontologies were about real world entities, while on the other hand, the actual content seemed to focus more on coding systems than real world entities. We confirmed that our interest is on coding systems aiming at classifying real world entities and addressed this remark by changing names of ontologies and classes and by adding appropriate explanations to this document.

We also received questions about the relations between our work of reengineering RTMS data into ontologies and other relevant work carried out within NeOn, especially concerning methodologies and design patterns. In this revision of the ontologies we highlighted where the connections are (see for example discussion in Sec. 4.1.3).

A related concern was about the decision to model some entities as instances instead of classes, as it would have been more intuitive. This is, for example, the case of the taxonomic classification of living organisms. This comment made us understand the need for better explanation of the domain at hand, the purpose of the RTMS and of the ontologies based on it. We addressed this comment by expanding the relevant sections in this document (cf. Sections 2.1, 2.3, 3.1, 4.1.1, 4.1.2, and 4.1.3).

Finally, the comment concerning the need to highlight the use case for networked/modularized ontologies was addressed by adding a discussion on the issue in Chapters 1 and 5.

[35] http://www.fao.org/countryprofiles/geoinfo.asp?lang=en

[36] http://www.fao.org/countryprofiles/

[37] http://www.neon-toolkit.org/wiki/index.php/OWLDoc

3.2.3 Ontology design

It was stressed that a more flexible design would separate the model of the ontologies from their population (as A-boxes importing their T-box). We took on board this comment, as can be seen in the network presented in this deliverable (Sec. 3.6).

Questions were raised concerning the reason to name individuals after their classification code, as opposed to using names in natural language (for example, in English). This comment is very sensible, but in many cases names in natural language are not available (cf. divisions of water areas) and even if they are, the choice of one language over the others (the six FAO languages, Arabic, Chinese, English, French, Russian, Spanish have equal importance) would be completely arbitrary. For these reasons, we kept the same names as in the previous deliverable, and adopted a uniform modeling and named instances with a concatenation of meta code and ID (prefixed with the string "ID"). Names in natural languages are represented as values of data type properties.

As an alternative, one could arbitrarily choose one or two languages (and, say, one type of names between "short" and "long" and "official") and use the `rdfs:label` to represent them. Unfortunately, the tool used to reengineer relational data into ontologies currently does not support this, so the addition of RDFS labels can only be done afterwards, by post-processing the data file.

It was pointed out that domains and ranges of some ontologies included in D7.2.2 could have been further specified (personalized) for each object property. In some cases, that was an intended style, meant to leave the modeling more general and less ontologically loaded (using general domain/range together with class restrictions), however we also acknowledge the fact that in other cases domains and ranges could have been quite appropriately personalized. We then accommodated these comments (cf. species ontologies, now specialized as a dedicated ontology to taxonomic classification of biological entities - which however still suffers from non-optimal modeling style, as discussed in Sec. 4.1.3).

3.2.4 Ontology implementation

We found very appropriate the remark about the lack of comments included in the ontologies released in D7.2.2. The cause for this was that we relied on the detailed explanations contained in the deliverable accompanying the ontologies, however we acknowledge the fact that it is quicker and safer if the most important comments and explanations are included in the ontologies themselves. So this is what we did this time, together with clear references to the detailed documentation included in this deliverable.

The ontologies released in D7.2.2 turned out to be mistakenly "duplicated", with consequent problems for their URIs. What happened there is that, due to some limitations - now overcome - of the NTK, it was very problematic to load the ontology on biological species, as that was too big. So we also distributed a model-only ontology (T-box only), in order to help the user and, finding this useful, we did the same for other ontologies. Unfortunately, URI's were not changed because only the T-boxes together with their corresponding A-boxes were considered to be the real outcome of our work. In this deliverable, we have brought forward our pre-mature attempt at modularization by separating T-box and A-box, taking care that URIs are clearly distinguished, and adding explanations wherever they are needed.

3.3 Reengineering of ASFA thesaurus

The ASFA Thesaurus[38] is an indexing and searching tool developed and maintained by the Aquatic Sciences and Fisheries Abstracts (ASFA) Partnership[39] to index the records contained in the ASFA

[38] http://www.fao.org/fishery/asfa/8/en

bibliographic database. ASFA abstracts cover the world's literature on fisheries (in particular, science, technology, management, and conservation of marine, brackish water, and freshwater resources and environments, including their socio-economic and legal aspects).

ASFA graph currently consists of more than 10,000 nodes. It has a typical thesaurus structure, made up of descriptors (graph nodes), equivalent terms, and relations among descriptors (edges between nodes, belonging to the following types: “broader term” (BT), “narrower term” (NT), “related term” (RT), “used for” (UF)) that create an indirect acyclic graph. Descriptors are directly encoded via a "preferred" English term.

The interest in reengineering the ASFA thesaurus as an ontology predates the NeOn project (see Sec. 3.3.2) and is motivated by the expected advantage in maintenance and exploitation resulting from having a resources with clear semantics and expressed by means of standard languages and technologies. From the point of view of the objectives of WP7, the main expected advantage of including ASFA in the fisheries network of ontologies is to widen the search possibilities over fisheries related resources, so as to include also search on textual repositories such as the ASFA collection of abstracts. Moreover, both the ASFA partnership and FAO[40] could profit from the alignment with the many standard classification already included in the network, both for the purpose of indexing and searching, and from the exploitation of the NTK and related plugins for all activities involved in the lifecycle of ASFA.

After a recap of the possible design patters to reengineer thesauri (Sec. 3.3.1), we present a reengineering of ASFA (Sec. 3.3.4) that builds on the lessons learned from previous attempts (Sec. 3.3.2) and on the requirements gathered within WP7 (Sec. 3.3.3). We also present a sample linking of the reengineered ASFA with the ontologies based on RTMS (Sec. 3.3.5).

3.3.1 Design patterns for reengineering Knowledge Organization Systems (KOS)

Thesauri are examples of Knowledge Organizations Systems (KOSes), also called Concept Schemes in SKOS (Simple Knowledge Organization Systems) [MB05]. In [D2.5.1], two design patterns are provided for KOS reengineering (see Figure 9):

(1) the KOS node and edge *types* (i.e. its schema) are converted with the semantics of a description logic T-box, and the KOS nodes and edges (*tokens*) with the semantics of a description logic A-box. E.g. `asfa:Descriptor` is converted into `[asfa:Descriptor rdf:type owl:Class]`, and `asfa:Dredging` is converted into `[asfa:Dredging rdf:type asfa:Descriptor]`.

(2) the KOS node and edge types are converted with the semantics of OWL metamodel, and the KOS nodes and edges with the semantics of a description logic T-box. E.g. `asfa:Descriptor` is converted into `[asfa:Descriptor owl:equivalentTo owl:Class]`, and `asfa:Dredging` is converted into `[asfa:Dredging rdf:type owl:Class ]`.

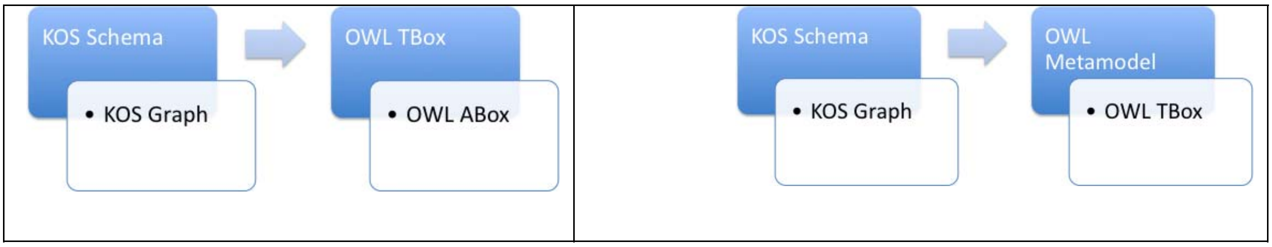


Figure 9. KOS reengineering pattern: A-box (left), T-box (right).

[39] http://www.fao.org/fishery/asfa

[40] FAO hosts the secretariat of the ASFA partnership.

The A-box pattern leaves the informal semantics of the reengineered resources mostly “untouched”. On the contrary, the T-box pattern tries to enforce a formal semantics to them, even at the cost of changing their structure. The following examples illustrate the difference:

- `(a1) asfa:Descriptor rdf:type owl:Class`
- `(a2) asfa:Dredging rdf:type asfa:Descriptor`
- `(t1) asfa:Descriptor owl:equivalentTo owl:Class`
- `(t2) asfa:Dredging rdf:type owl:Class`

In (a1), the A-box reengineering pattern is applied to ASFA schema: a mapping axiom is added which states that `asfa:Descriptor` is an instance of `owl:Class`. In (a2), a particular `asfa:Descriptor, asfa:Dredging`, is said to be an instance of `asfa:Descriptor`. In formal semantic terms, ASFA schema has been considered as the “vocabulary” of ASFA, while its nodes as the “domain” of ASFA.

In (t1), the T-box reengineering pattern is applied to ASFA schema: a mapping axiom is added which states that `asfa:Descriptor` is equivalent to `owl:Class`. In (t2), a particular `asfa:Descriptor, asfa:Dredging`, is said to be an instance of `owl:Class`. In formal semantic terms, ASFA schema has been aligned to the OWL “metamodel”, and ASFA nodes have been considered as the “vocabulary” of ASFA. The “domain” of ASFA is considered in this case as the interpretation of ASFA nodes as formal classes: e.g. Dredging being interpreted as the set of all dredging activities actually occurring in the fishery world.

It is straightforward to see that the interpretation exemplified by the A-box pattern is actually the interpretation of the KOS, not of the fishery world: as such, it does not imply any (even plausible) change in the original semantics. With the T-box pattern, things are more troublesome: how can we be sure that fishery experts actually intend something that corresponds to a formal (extensional) interpretation of ASFA nodes as classes of real world activities? With the T-box pattern, we are *changing the semantics of the KOS*. This change can be cost-effective or not. While it is beneficial to have a rigorous extensional semantics associated with ontologies, pragmatically it can be the opposite: experiences in FOS [GFK+04] and NeOn, indicate that when a large organization has a lot of conceptual structures already represented and maintained as KOSes, it is technically and socially difficult to impose a T-box reengineering pattern. On the contrary, the broad success of SKOS indicates the practical advantage of the A-box reengineering pattern.

SKOS (Simple Knowledge Organization Systems, [MB05]), is the most used metamodel to reengineering thesauri as OWL or RDF(S) models. SKOS has been thought primarily as an RDF(S) vocabulary, because it is intended to represent thesaurus nodes as RDF descriptions (called “concepts”), which are on their turn asserted to be rdf:type rdfs:Class. This ambiguity cannot be expressed in OWL-DL, thus preventing the full capability of OWL-DL reasoners such as Fact++[41] or Pellet[42]. Representing KOSes in SKOS can be considered as one variety of the A-box KOS reengineering pattern. An example of SKOS conversion for ASFA is the following:

- `(s1) asfa:Descriptor rdfs:subClassOf skos:Concept`
- `(s2) skos:Concept rdf:type rdfs:Class`
- `(s3) asfa:Dredging rdf:type asfa:Descriptor`

It is basically an A-box pattern, but the intermediate class `skos:Concept` introduces a novelty: `asfa:Descriptor` is no more `rdf:type owl:Class`, as in (a1), but `rdfs:subClassOf skos:Concept`, and (s2) `skos:Concept rdf:type rdfs:Class`. Therefore, `asfa:Descriptor rdf:type rdfs:Class` (because of inheritance). `rdfs:Class`, differently from `owl:Class`, is not constrained to one specific

[41] http://owl.man.ac.uk/factplus/

[42] http://pellet.owldl.com

logical layer, and can be either a class or a metaclass. SKOS reengineering pattern can therefore be used to say consistently the following axioms:

- `(s4) asfa:Dredging rdf:type asfa:Descriptor` *(as in (a2))*
- `(s5) asfa:Dredging rdf:type rdfs:Class`

(even more generally than in (t2), since owl:Class rdfs:subClassOf rdfs:Class)

In other words, SKOS reengineering pattern allows *in principle* to reengineer KOSes with both A-box and T-box patterns. However, in practice this is not really true, since it generates an OWL-Full ontology, which limits the applications of reasoning engines for OWL-DL.

On the reasoning side, we should notice that the A-box pattern is obviously weaker, since the BT relation alone, even if reengineered as an `owl:TransitiveProperty` as in `skos:broaderGeneric`, does not ensure the powerful inheritance reasoning available with the T-box pattern. Recent developments of HiDL-Lite [HDL] reasoning promise to allow for the A-box pattern, without requiring the SKOS “two-layered” pattern.

A further hybrid pattern that makes use of a more complex reengineering procedure is the following: firstly (A-box reengineering pattern), a KOS is A-box-reengineered by aligning it to a metamodel called Linguistic Metamodel (LMM), with explicit mappings to possible formal interpretations; secondly (T-box interpretation pattern), it can be reengineered as a T-box by declaring what mappings to what formal interpretations are required.

3.3.2 Previous ASFA reengineering effort

Previous studies aimed at reengineering ASFA include the FOS project ([GFK+04], also reported in [D7.2.2]), and reengineering done within the oneFish[43] and the AquaRing[44] projects.

Within the FOS project, a T-box reengineering pattern was used, leading to a DAML+OIL ontology (later migrated to OWL) of 11484 classes, where the BT relation was simply reengineered as a `rdfs:subClassOf` relation, for example:

- `(t3) asfa:Decompression_chambers rdfs:subClassOf asfa:Diving_equipment`

sometimes this leads to counterintuitive axioms, e.g.:

- `(t4) asfa:Blood_vessels rdfs:subClassOf asfa:Circulatory_system`

In (t4), the extensional semantics resulting from the reengineering leads to the counterintuitive interpretation that all blood vessels *are* circulatory systems, while the intended meaning of ASFA experts was probably *intensional*, i.e. that blood vessels (as a class of actual blood vessels) are *part of the* circulatory system (as a class of actual circulatory systems).

(t4) axiom also shows an example of ambiguity arising with this kind of reengineering: `asfa:Blood_vessels` has to interpreted as the class of individual blood vessels (since plural names in thesauri do not necessarily refer to collection entities), or as the class of collections of blood vessels in an organism? The T-box pattern forces to take one of these interpretations, while the A-box pattern does not, because its domain of interpretation is constituted by ASFA terms, not of actual entities from the fisheries world.

In the FOS project, after ASFA T-box reengineering, about 1600 ASFA descriptors that had no BT relation (“orphans”) have been semi-automatically aligned to external ontologies: other fishery ontologies when possible (e.g. fos-economy, fos-techniques), otherwise OntoWordNet (with prefix `own` in the example) and DOLCE (with prefix `dul` in the example), and validated with the help of a fishery expert, Ian Pettman (see below). Some examples include:

[43] http://www.onefish.org

[44] http://www.project.aquaringweb.eu/index.html

- `(t6) asfa:Agropisciculture rdfs:subClassOf fostechniques:Aquaculture`
- `(t7) asfa:Fish:utilization rdfs:subClassOf foseconomy:PostHarvestUse`
- `(t8) asfa:Entrainment rdfs:subClassOf own:Movement__Motion`
- `(t9) asfa:Energy_transfer rdfs:subClassOf dul:Process`

The problem of orphans is relevant also in A-box-reengineered versions. In the current version, there are still 1482 orphan descriptors.

In 2008, Ian Pettman from OneFish produced two new OWL versions of ASFA, by performing firstly a T-box-reengineering, and afterwords a SKOS alignment. Since this version of ASFA is the last and the most correct one from the ASFA partnership, it has been used to produce the NeOn version (see Sec. 3.3.4).

Pettman's version was created also on demand from the EU Aquaring project. In this version, descriptors are taken as instances of the Descriptor class; the BT, RT, USE were transformed in properties: hasParent, related_to and synonym. However, it was found that BT and RT cannot be always transformed to hasParent and related_to concept idea. They extracted 1715 parent terms with no BT, and started aligning them under ASFIS subject categories classification elements, but they gave up after 139 alignments. The purpose of Aquaring is to use ASFA ontology (among other ontologies) to semantically annotate resources. They had problems with the multilingual support compatibility with the other ontologies they have.

3.3.3 Current requirement for ASFA reengineering

The following core requirements for the reengineering of ASFA were gathered by FIES as part of T7.2, based on interviews with the ASFA team and on FAO needs for interfacing with the ASFA thesaurus. ASFA thesaurus is currently used a) to index ASFA records with standard keywords, and b) to search for ASFA records with standard keywords.

Maintenance is a key concern, and in fact ASFA is looking for a new way in which to manage the thesaurus, so that it can be updated more frequently (an ontological OWL format could be a choice). In particular, since maintenance is ensured by editors (approximately six people) belonging to different organizations, geographically distributed, editing rights, user roles (e.g. editor, validator), concurrent editing should be managed accordingly. Digital storage/backup are needed, together with web-based access and versioning system. Also, a mechanism to manage test and production versions (and to replace the former with the latter when needed) should be in place.

As for **accessibility**, the thesaurus is currently distributed in print and also accessible via different web interfaces. These formats will continue to be needed.

As for the **modelling** to adopt, the thesaurus needs to maintain at a minimum its current expressivity as found in Ian Pettman's OWL version: use, used for, status, audit dates, broader term, narrower term, related term. The modelling adopted should support multilinguality, and it should be comprehensible to anyone with an indexing background.

It should be possible to perform **integration** of the thesaurus with other resources, such as AGROVOC[45], the corporate FAO thesaurus for document indexing.

In terms of **performance**, one should be able to open and edit the ontology on a standard specification Intel computer (Pentium 4, 2.00 Ghz, 1gb RAM) without significant delay (ontology opened in 0-5 seconds).

The reengineered thesaurus should not be dependent on outside data sources or models except where these are **standard**, e.g. SKOS. ASFA thesaurus is modelled after ISO 2788[46], it would be

[45] http://www.fao.org/agrovoc/

good that the thesaurus is in or can be converted to a standard model, e.g. SKOS[47]. Syntactically, the reengineered thesaurus should be expressed in a standard markup language, such as OWL or RDF.

It is desirable that the thesaurus can be easily split into thematic "**modules**" (such as Biological Sciences and Living Resources, Ocean Technology, Policy and Non-Living Resources, Aquatic Pollution and Environmental Quality, Aquaculture, Marine Biotechnology).

3.3.4 ASFA reengineering

Lessons learnt and requirements (e.g. efficiency, standards) seem to go to the direction of an A-box reengineering, with an alignment to SKOS as described below. Two versions of ASFA A-box KOS reengineering pattern have been produced. The first one adopts the pure A-box pattern, the second aligns the first to SKOS. Further alignment to LMM is provided by the skos2lmm module (see Annex VI). All versions are based on two namespaces:

`http://www.xxx.org/asfam.owl, and http://`*`www.xxx.org`*`/asfad.owl`

where `www.xxx.org` is a placeholder for the namespace that will be indicated by ASFA partnership. `asfam.owl` is the file with the ASFA schema in OWL; `asfad.owl` is the file with the complete ASFA data (descriptors, non-descriptors, and relations between them).

The current versions are downloadable from the following CNR URIs:

1. ASFA schema: http://www.ontologydesignpatterns.org/ont/fao/asfa/asfam.owl
2. ASFA instances: http://www.ontologydesignpatterns.org/ont/fao/asfa/asfad.owl
3. aligned ASFA schema to SKOS:

   http://www.ontologydesignpatterns.org/ont/fao/asfa/asfam2skos.owl
4. alignments of ASFA schema to LMM:

   http://www.ontologydesignpatterns.org/ont/fao/asfa/asfam2lmm.owl

The following `owl:imports` axioms hold: `asfad.owl owl:imports asfam.owl; asfam2skos.owl owl:imports asfam.owl; asfam2lmm.owl owl:imports asfam2skos.owl`. Changes with versions until now, issues arisen, and things to do are all annotated in the annotation spaces of the ontologies. The ASFA schema has been fully reengineered in OWL as follows (see Figure 10):

- ASFATerms has been represented as:
  - `asfam:ASFATerms rdf:type owl:Class`
  - `asfam:ASFATerms rdfs:subClass skos:Concept`
- ASFA Descriptor has been represented as:
  - `asfam:Descriptor rdf:type owl:Class`
  - `asfam:Descriptor rdfs:subClass asfam:ASFATerms`
- ASFA NonDescriptor similarly as ASFA Descriptor
- BT has been represented as:
  - `asfam:BT rdf:type owl:TransitiveProperty`
  - `asfam:BT rdfs:subPropertyOf skos:broader`
- NT has been represented as:

[46] http://www.collectionscanada.gc.ca/iso/tc46sc9/standard/2788e.htm

[47] http://www.w3.org/2004/02/skos/

  - `asfam:NT rdf:type owl:TransitiveProperty`
  - `asfam:NT rdfs:subPropertyOf skos:narrower`
  - `asfam:NT owl:inverseOf asfam:BT`
- relatedTerm has been represented as:
  - `asfam:relatedTerm rdf:type owl:SymmetricProperty`
  - `asfam:relatedTerm rdfs:subPropertyOf skos:related`
- USE has been represented as:
  - `asfam:use rdf:type owl:ObjectProperty`
- USED FOR has been represented as:
  - `asfam:usedFor rdf:type owl:ObjectProperty`
  - `asfam:usedFor owl:inverseOf asfam:use`

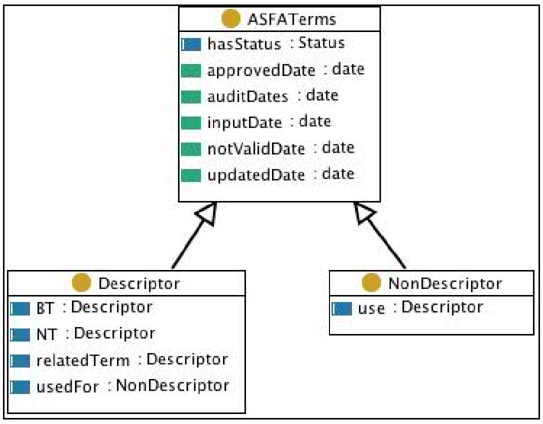


Figure 10. The ASFATerms class and subclasses in asfam.owl.

Additional attributes for term management, such as input date, approved date, audit date, etc. have been represented as instances of `owl:DatatypeProperty` on the class `asfam:ASFATerms`.

The alignment of the OWL version of ASFA to SKOS has been made straightforwardly by declaring the following axioms (cf. Figure 11 below):

- `asfam:ASFATerms rdfs:subClassOf skos:Concept`
- `asfam:BT rdfs:subPropertyOf skos:broader`
- `asfam:NT rdfs:subPropertyOf skos:narrower`
- `asfam:relatedTerm rdfs:subPropertyOf skos:related`
- `asfam:use rdfs:subPropertyOf skosmapping:exactMatch`
- `asfam:usedFor rdfs:subPropertyOf skosmapping:exactMatch`

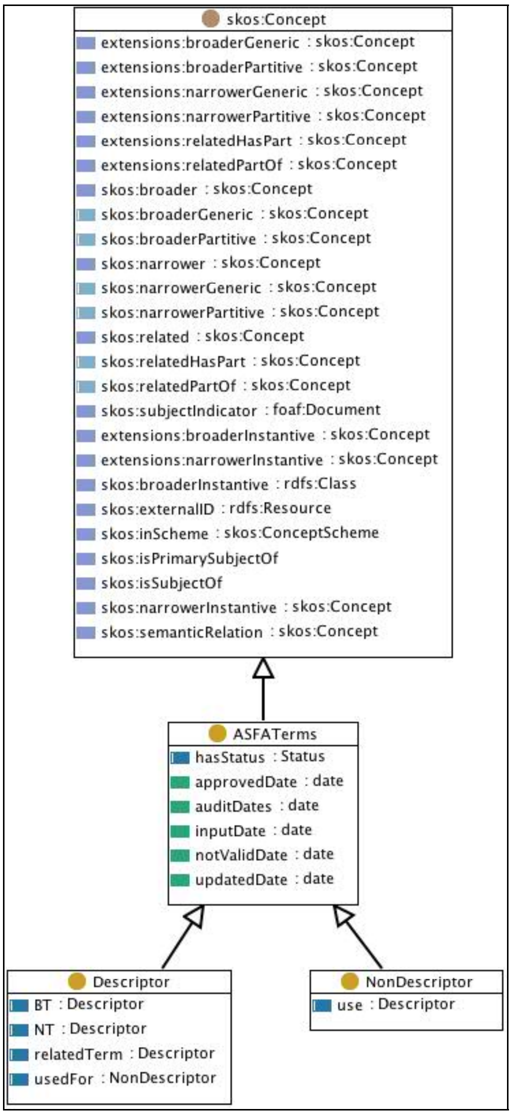

skos:Concept
extensions:broaderGeneric : skos:Concept
extensions:broaderPartitive : skos:Concept
extensions:narrowerGeneric : skos:Concept
extensions:narrowerPartitive : skos:Concept
extensions:relatedHasPart : skos:Concept
extensions:relatedPartOf : skos:Concept
skos:broader : skos:Concept
skos:broaderGeneric : skos:Concept
skos:broaderPartitive : skos:Concept
skos:narrower : skos:Concept
skos:narrowerGeneric : skos:Concept
skos:narrowerPartitive : skos:Concept
skos:related : skos:Concept
skos:relatedHasPart : skos:Concept
skos:relatedPartOf : skos:Concept
skos:subjectIndicator : foaf:Document
extensions:broaderInstantive : skos:Concept
extensions:narrowerInstantive : skos:Concept
skos:broaderInstantive : rdfs:Class
skos:externalID : rdfs:Resource
skos:inScheme : skos:ConceptScheme
skos:isPrimarySubjectOf
skos:isSubjectOf
skos:narrowerInstantive : skos:Concept
skos:semanticRelation : skos:Concept
ASFATerms
hasStatus : Status
approvedDate : date
auditDates : date
inputDate : date
notValidDate : date
updatedDate : date
Descriptor
BT : Descriptor
NT : Descriptor
relatedTerm : Descriptor
usedFor : NonDescriptor
NonDescriptor
use : Descriptor

Figure 11. The ASFATerms class and subclasses in asfam.owl as aligned to SKOS.

3.3.5 Networking ASFA and the RTMS-based ontologies

The main problem that one encounters when mapping `asfad.owl` to RTMS_based ontologies is the *heterogeneity* of their semantics. RTMS-based ontologies have different domains of interpretations: e.g. *water areas, taxa* (organism types), *commodities* (fishery product types), *vessel classifications* (vessel types), etc., while `asfad.owl`, as reengineered with the A-box pattern, has only one domain of interpretation: the ASFATerms (either descriptors or non-descriptors), whose proper real world interpretation is left to the user.

The result of interpretation heterogeneity is that links between ASFA and RTMS-based ontologies must adopt different *correspondence patterns* [D2.5.1] according to the semantics of the two resources. Correspondence patterns describe *types* of mapping, e.g. *class is subclass of class* (`?owl:Class rdfs:subClassOf ?owl:Class`), *class is equivalent to class* (`?owl:Class owl:equivalentClass ?owl:Class`), *individual has type class* (`?rdf:Description rdf:type ?owl:Class`), *individual is subclass of class* (`?rdf:Description owl:equivalentClass ?owl:Class`), etc. In the following examples, encoded in the ontology:

`http://www.ontologydesignpatterns.org/ont/fao/asfa/asfad2rtms.owl` some mappings proposed by the fishery expert Ian Pettman in 2003[48] are used in order to sample the mapping from ASFA to RTMS ontologies:

– using the pattern **individual is equivalent to class**:

`asfad:Aquatic_organisms` has most `asfam:NT` axioms that describe informal, non-Linnean types of aquatic organisms; it is therefore advisable to map it equivalently to the non-Linnean taxonomy from RTMS, i.e. the commodity classification[49]:

```
(m1) asfad:Aquatic_organisms
     owl:equivalentClass FI_commodities:FI_commodity_classification
```

`asfad:Stocks` has its `asfam:NT` axioms that describe the same kind of entities as the ones classified under the class `stock:Stock` from RTMS,[50] therefore we can reasonably map them equivalently:

```
(m2) asfad:Stocks owl:equivalentClass stock:Stock.
```

`asfad:Water_bodies` has its `asfam:NT` axioms that describe the same kind of entities as the ones classified under the class `water:FAO_fishing_area` from RTMS[51], therefore we can reasonably map them with an `owl:equivalentClass` axiom:

```
(m3) asfad:Water_bodies owl:equivalentClass water:FAO_fishing_area
```

– using the pattern **individual is subclass of class**:

`asfad:Ocean_space` is described in way compatible to the class `water:FAO_fishing_area` from RTMS, although it is explicitly bound to “legal aspects”, therefore we can reasonably map them with a `rdfs:subClassOf` subsumption:

```
(m4) asfad:Ocean_space rdfs:subClassOf water:FAO_fishing_area
```

Interestingly enough, `asfad:Ocean_space` has no `asfam:BT` axioms to `asfad:Water_bodies`, probably because it was considered for a legal usage, disjoint from typical usages of water area identification.

– using the pattern **individual has type class**:

`asfad:Research_vessels` (like the next four examples provided in (m6) to (m9)), is described in way compatible to the class `vessels:vessel_classification` from RTMS,[52] therefore we can reasonably map them with a `rdf:type` axiom:

- `(m5) asfad:Research_vessels rdf:type vessels:vessel_classification`
- `(m6) asfad:Protection_vessels rdf:type vessels:vessel_classification`
- `(m7) asfad:Drilling_vessels rdf:type vessels:vessel_classification`
- `(m8) asfad:Survey_vessels rdf:type vessels:vessel_classification`
- `(m9) asfad:Fishing_vessels rdf:type vessels:vessel_classification`

The sample proposed here, based on expert’s validation, show the formal problems emerging from heterogeneous resource mappings: some correspondence patterns are *interpretation-preserving*, i.e. the semantics of the two resources match without formal problems, while other patterns are not, and therefore induce an OWL-Full semantics.

[48] At that time, a different set of RTMS-based ontologies was used, but the differences with the current ones are not relevant to this mapping task.

[49] http://www.fao.org/aims/aos/fi/FI_commodities_v1.3.owl#FI_commodity_classification

[50] http://www.fao.org/aims/aos/fi/stock#Stock

[51] http://www.fao.org/aims/aos/fi/water#FAO_fishing_area

[52] http://www.fao.org/aims/aos/fi/vessels#vessel_classification

The first case occurs with mappings (m5) to (m9), which adopt the *individual has type class* pattern, while the latter case occurs with mappings (m1) to (m4), which adopt the *individual is equivalent to class* and the *individual is subclass of class* patterns.

However, the non-interpretation-preserving mappings match the ASFA version with alignment to SKOS, which is already in the OWL-Full semantics.

3.3.6 Sample modularization of `asfad.owl`

Once mappings are available, ASFA can be given some modularization on the basis of the RTMS architecture distinguishing between species, resources, vessel classifications, gear classifications, techniques, geographic entities, etc. As an example, let's consider the (m2) mapping: this suggests a meaningful module that can be extracted from `asfad2rtms.owl` by leveraging e.g. the graph induced by `asfam:NT` and `asfam:relatedTerm` axioms involving `asfad:Stocks`. In Figure 12, an example of such a module is shown. The module, encoded in the ontology `asfastocks.owl`[53] is obtained with the application of the following SPARQL CONSTRUCT query run against the mapping ontology: `asfad2rtms.owl:`[54]

```
CONSTRUCT { ?x ?r ?y }

WHERE {

  ?x ?r ?y .
  ?x ?m stock:Stock }
UNION {
  ?x ?r ?y .
  ?z ?s ?x .
  ?z ?m stock:Stock }
}
```

The module shown in Figure 12 is extracted with a SPARQL CONSTRUCT from `asfad.owl`, after a mapping from `asfad:Stocks` to RTMS' `stock:Stock`. The depth of the graph is kept to two levels for visualization purposes, but the CONSTRUCT can be designed in order to obtain deeper modules (as symbolized by the arrows on the right).

[53] http://www.ontologydesignpatterns.org/ont/fao/asfa/asfastocksT-box.owl

[54] http://www.ontologydesignpatterns.org/ont/fao/asfa/asfad2rtms.owl

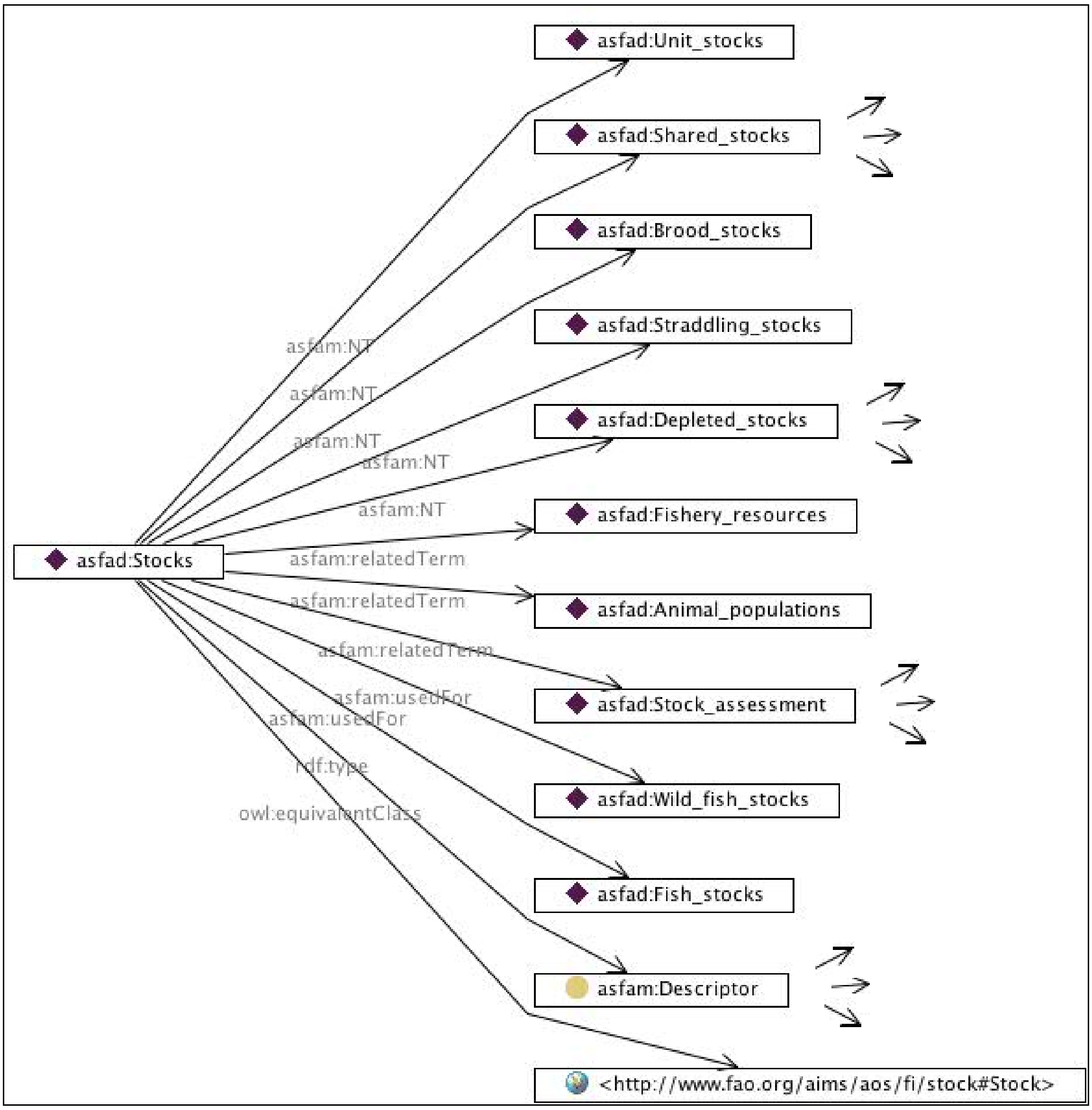


Figure 12. An excerpt from `asfaT-box.owl`.

The ontology module so created can then be transformed automatically into a T-box ontology, e.g. with the purpose of running more complex inferential tasks.

3.4 Reengineering the reference tables as a network of ontologies

In this section we describe the process of reengineering RTMS into individual ontologies (Sec. 3.4.1) and into networked ontologies (Sec. 3.4.2).

3.4.1 Reengineering individual ontologies

Given the implementation of the RTMS, we found technically convenient, and compatible with common understanding and explanation of fisheries domain, to use (fragments of) the hierarchy of types to create reengineered ontologies. This approach, together with a highly modular modelling style, resulted in specialized ontologies for each coding system. Since the data at hand is stored in relational form the process of converting it into ontologies is at the same time a problem of domain modelling and data reengineering. This was achieved by first following several iterations of domain conceptualization and actual population of the ontologies. Then, the resulting stable models underwent assessment by CNR focusing the ontology modeling and the feedback so gathered was applied in modeling and population iteration. In some cases, technical limitation or pragmatical considerations prevented the immediate application of some suggestions, and contributed to the development of the discussion presented in Chapter 0. As ontology editor, we used NTK 1.2.1, Protégé 3.3.1 and Protégé 4.

In order to obtain an adequate knowledge of the domain covered by the reference data, we studied all the available material, including the relevant fact sheets from the Handbook of Fishery Statistical Standards by the Coordinating Working Party on Fishery Statistics (CWP) and the classification systems used, as published by the organizations that maintain them. After having obtained a general overview of the domain, we interviewed domain experts who gave us a practical understanding of the rationale behind the adopted classification systems and of the connections between the reference data and the statistical data collected. Moreover, we also had a number of interviews with the information experts working with the RTMS. From these interviews we obtained a deeper understanding of the database and of the modelling choices it implements. They also gave us insight into the lifecycle of the reference data in the context of real applications and actual use.

The tool we used to access and reengineer the data in the DB is ODEMapster[55] [BAR06, BAR07], developed at the Universidad Politecnica de Madrid and currently integrated in the NTK as a plug-in.[56] To date, ODEMapster is the only NeOn tool that allows one to lift relational data according to OWL ontologies (NTK also includes a functionality to lift relational databases to ontologies, working with F-logic ontologies). ODEMapster was already used during the work that led to D.7.2.2: chapter 7 in that deliverable contained detailed feedback regarding its functionalities and limitations.

ODEMapster is an engine that executes mappings between an ontology model and a database by means of a declarative language, R2O. R2O allows the description of mapping expressions between ontology elements (concepts, attributes and relations) and relational elements (relations and attributes). It is based on conditions and operations and on rule-style mapping definition for attributes. R2O is independent from the particular RDBMS used. The ODEMapster processor generates ontology instances from relational instances based on the mapping description expressed in an R2O document. It can operate at run-time (on-demand query translation) or it can perform massive batch processes that generate all possible ontology individuals from the data repository. The processor delegates the execution of certain actions to the RDBMS and executes the rest by itself (post processing). The main steps of its executions are: Query and R2O parsing, SQL generation, RDBMS execution result grouping and finally post-processing.

Since D7.2.2, ODEMapster has improved in the following respects: it is now integrated as a plug-in of the NTK; it has improved user interface, through which it is now possible to access both MySQL and Oracle; it supports self-joins. However, we found that the user documentation is still to be improved, as well as for the error and warning messages. Also, a mechanism to automatically change URI in the entire R2O file is still missing, which forces the user to manually change (search & replace) all URIs every time it is needed (for new versions, for example). Detailed feedback on ODEMapster is provided in Sec. 4.1.11.

[55] http://www.neon-toolkit.org/wiki/index.php/ODEMapster

[56] However, at the time of the work reported in this document, it was not integrated with the latest versions of NTK (NTK 1.2.1 and NTK 1.2.2). All comments reported in this document refer to the ODEMapster engine as used as a stand-alone tool, from command-line, and applied to a MySQL database.

3.4.2 Reengineering networked ontologies

The network of ontologies makes extensive use of `owl:imports` statements, as they are used to connect T-box and A-box, and to define object properties having domain and range defined in different ontologies. This latter operation is supported by NTK (note that NTK assumes that the imported ontology is loaded before the one containing the imports statement), as well as by other common tools such as Protégé 3.3 and Protégé 4.

From the fact that fisheries ontologies are designed to have no overlaps in their covered domains, and all the data is modelled as instances, it follows that all actual links between ontologies are at the A-box level. This implies that we devote careful consideration to at least three aspects: where data about linking can be found and how it can be extracted, how links can be represented, how links can be processed. In this section, we focus on the first point, i.e. where link data is stored and how it can be used. We distinguish the following three cases.

1) Linking data is in RTMS

In some cases, the RTMS contains information about correspondences between reference data (e.g. between ISSCAAP classification and taxonomic classification of species). Therefore this information is available when reengineering the corresponding individual ontology. In general, this happens every time that data is used for statistical data collection, and the connection between data is established by international agreements or conventions (e.g. membership in organizations, correspondences between commodities, jurisdiction on land and water areas), or results from specific analysis of the data conducted by FAO or third bodies (e.g. overlapping between water areas, often done by means of GIS technologies).

Given the modelling adopted, the links extracted from the RTMS correspond to links between ontologies at the instance level. As for the extraction of the linking data, we identify 4 main steps:

1. identify theoretical connections between pieces of data;
2. create an appropriate ontology schema (T-box) or edit an existing one. In particular: add appropriate object properties with domain/range specification. For the extraction of the data and implementation of the links, it is important to pay attention to the "direction" of linking (e.g. ISSCAAP group contains species, or species belongs to ISSCAAP group, or both) and to the direction of the import statement;
3. analyze DB and manipulate it as needed (usually: creation of views for data normalization);
4. generate instance file (A-box) with the available tool.

The last step was achieved with ODEMapster used in batch mode (for the twofold reason that we had to manipulate a database on which we have no ownership, and in order to extract instances that could be exposed to the NeOn partners and to the general public). ODEMapster generates an RDF file with no header, nor import statement. However, the main limitation of that tool is that the modularity of the ontology design is not reflected in the r20, which is rather monolithic. We further discuss this point in Sec. 4.1.11.

Linked instances are represented by their URIs.

2) Linking data is in semi-structured documents

RTMS may not contain all pieces of information needed to link ontologies. A typical example is the connection between water areas and biological entities, which "form" the concept of stock (under the operational definition that a stock is a population of one or more biological entities living in one or more water areas).

In this case, the linking information is to be found elsewhere, in corpora of semi-structured data, i.e. fact sheets. For stocks, good sources of information are the fact sheets about aquatic resources which are published by FAO and based on data provided by Regional Fisheries Bodies[57] (RFB) specific for the given stock. Note that since RFB focus on specific "regions" or portions of water areas, they adopt a variety of classification of water areas (i.e. this implies gathering data about "conversion" between water systems as well).

Given the modelling of the ontologies, this is again a case of linking at the instance level. As for the extraction of the linking data, the 4-step process described above applies here as well, with the difference that "point 3) Analyze DB" is now replaced by the following:

a) parse the XML to extract the codes identifying water areas and species (e.g. FAO divisions, taxonomic code, specific water division as used by RFB);

b) if needed (and possible), apply conversion between classification systems;

c) extract from DB the information concerning the objects to be linked;

d) resolve ID from the DB as URI from ontologies;

e) export the data according to the language accepted by the tools.

Points d) and e) refer to the representation of the linking data. At this stage, we found convenient to store the data so extracted in new DB tables and to apply the tool for reengineering as in the previous case. However, this operation should be carefully organized with the owner of the database in order to make this process not simply a one-off, but a repeatable process.

3) Linking data can be "inferred" from data and/or ontologies

In some cases, links between entities in the ontologies are not established by international agreements and other similar agreements, and are not currently maintained in the RTMS or in other available sources. In such situations, connections may be "discovered". This is, for example the case for commodities and their "biological" source (the commodity "canned tuna" is obviously originated from "tuna"), or for gear types and the vessels that mount that gear.

In principle, the extraction of this type of information is rather trivial, as it may be as simple as a case of information extraction based on pattern matching. However, the whole process of establishing these links is not as trivial as it may seem. In fact it requires at least that:

1. aggregation, disambiguation and NLP analysis be implemented (e.g. "canned tuna" is used in commodity description, but is not sound from a biological point of view, since no species is called just "tuna");
2. the process be semi-automatic, i.e. fisheries experts should be able to revise/approve/discharge the information automatically found (this implies a GUI and a human computer interaction model suitable to the users involved);
3. the process be integrated into an ontology editing environment.

In this first version of the network, we have one case of this linking: between exclusive economic zones and countries, but this is necessarily going to play a larger role in the next release of the fisheries network. For a discussion on the representation and processing of this type of links, see Sec. 4.1.11.

[57] Regional Fisheries Bodies are international organizations whose members are countries, and whose mandate is to monitor the status of specific aquatic resources in a given water area. For a list of them, together with their area of competence, please refer to: http://www.fao.org/fishery/rfb/search/en#rfb_map (this link points to the world map of RFB).

4) Linking data is to be added manually

Finally, linking information may be stored in no data source, either because new (e.g. new membership in a RFB), or because it has changed over time (data is amended, new data is found), or because not directly accessible (e.g. contained in third party reports). In this case, linking data has to be generated manually by experts in the domain. Note that this case may be reduced to a general case of data maintenance, part of a normal editorial workflow.

Currently, it is possible to manually edit links between instances, both NTK and with Protégé, but the “integration” of data provided by human editors and data coming from other sources is an open issue which concerns at least data storage, versioning, and integrity. In fact, most of the data we dealt with comes from the RTMS, but we can only access and/or expose that data through ontologies.

Summarizing, the network presented in this document mainly deals with the first two cases (namely with linking data either present in the RTMS, or extracted from semi-structured corpora), but it also began experimentation with the third case.

3.5 The first network of fisheries ontologies

The first network of fisheries ontologies include ontologies about: taxonomic classification of biological entities, ISSCFC and HS classification of fisheries commodities, ISSCAAP classification (with links to taxonomic classification of biological entities and ISSCFC commodities), FAO division of water areas, large marine ecosystems (with links to FAO divisions), a “module” from the geopolitical ontology, exclusive economic zones (with links to geopolitical ontology), stocks (with links to taxonomic classification of biological entities and FAO divisions). We also include two ontologies that are not currently linked to other ontologies: the one dealing with the ISSCFG classification of gear types and the one dealing with the ISSCFV classification of vessel type and size (we plan on networking them in the next release). These two ontologies have been revised and improved with respect to the version presented in Deliverable D7.2.2, however they have not been included in the network because their linking corresponds to cases 3) and 4) as described in the previous section. We expect to be able to include them in the network (in particular, with links to: water areas, stocks) in the next release. Figure 13 below provides a graphical representation of the network based on RTMS: circles represent T-box, diamonds represent A-box. Arrows represent the direction of the `owl:imports` statements.

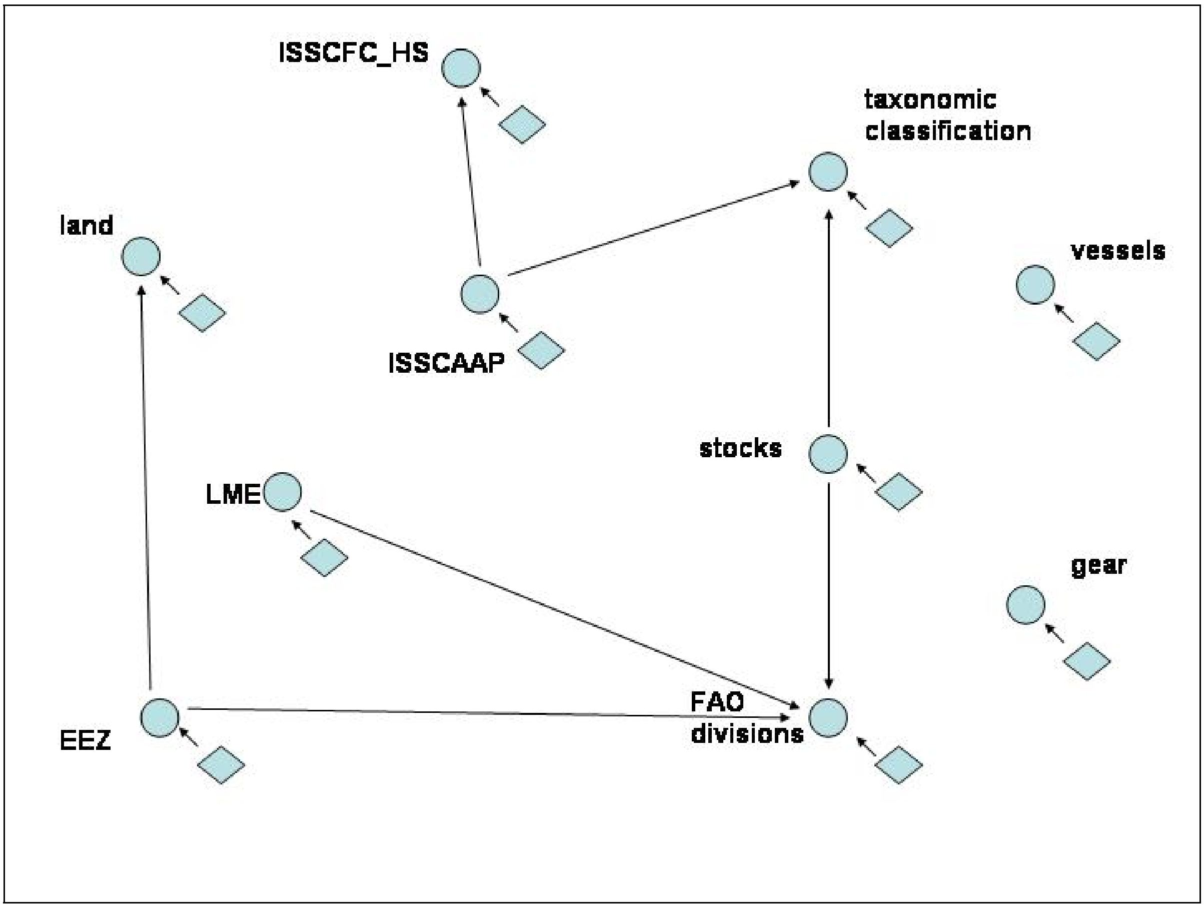


Figure 13. Sketch of the initial network of fisheries ontologies based on RTMS.

Contrary to what we did in D7.2.2 we do not include a detailed description of all ontologies in the deliverable. Instead, we use this document to describe the salient features of the network, discuss the issues that arose during the work, and account for problems encountered which should be addressed in the next revision of the network. Comments are added inside each ontology (all available at: http://www.fao.org/aims/neon.jsp) in order to allow anybody using the ontologies to understand their modelling and rationale. Moreover, the most relevant parts of this document are published together with the ontologies, and we also publish the OWLDoc (for NTK) version of each of them.

3.6 Features of the network

The approach followed is to keep T-box and A-box separated (in Figure 13 T-box is represented with a circle, while the A-box with a diamond), so that T-box can be used both for run-time access to the DB (or reused with other compatible data for different applications) and in conjunction with the data we provide. The data files import their corresponding ontology. Both A-box and T-box share the same namespace, but have different URIs. All files are serialized as RDF/XML.

The ontology URI is specified in the `xml:base` attribute, which is meant to correspond to the URL where the ontology is published. All ontological elements (classes, instances, properties) are defined in a dedicated namespace, one per domain: for example, the concept “stocks” is the defined in the namespace `http://www.fao.org/aims/fi/stock#` and it uses concepts defined in other namespaces (e.g. the concept "species" defined in `http://www.fao.org/aims/fi/taxonomic#` and the concept "FAOdivision" defined in `http://www.fao.org/aims/fi/water#`). Instances of a given class are then given the URI formed by the namespace where that class is defined, concatenated with the “meta” code and the ID as coming from the database. The idea behind this is to have classes and instances defined in the same namespace, even though they are actually contained in different ontologies/files, with different URIs. So far, namespaces have been kept the same through the various versions and revisions, so there is no notion of version embodied in the namespaces and in the URIs based on them. This may change in the future.

The `owl:imports` statements are always declared in the file with instances, also to allow one to inspect the T-box without having to load the instances (remind that NTK assumes that only ontologies loaded in the project may be imported; it also assumes that the imported ontology is loaded before the importing one).

When concepts are defined compositionally from concepts defined in other ontologies, the corresponding T-boxes are imported. A typical example is the concept “stock”, the definition of which would not be possible without referring to the concept of “species” and of “water area”. The stock ontology then imports the corresponding ontologies on taxonomic classification and FAO divisions of water areas (and the A-box on stocks imports the corresponding A-boxes on taxonomic classification and water areas).

Wherever possible, we tried to reuse ontology design patters available in the library of ontology design patters (cd. D2.5.1). For a discussion on the implementation of the ODP for taxonomic classification, see Sec. 4.1.3.

Names of ontology elements (classes, properties) are based on English, while names of instances are taken from a concatenation of their meta code (according to the hierarchies of types in the DB, cf. 3.1) and their ID from the DB. This is done in order to smooth the connection with the underlying database, since this is the practice commonly used in the DB to ensure uniqueness of the instance names (across the entire set of ontologies). The disadvantage of such an approach is that it is not well readable for humans. We considered also using the EN name for them, but we faced the problem that in many cases English names were missing (main example is the water areas). For a discussion on this issue, see Sec. 4.1.8. The list of naming conventions adopted is in Appendix I.

We also adopt a somewhat non-standard naming of datatype properties, in that besides using the `xml:lang` attribute we also add the two-character ISO code of the language in the name of the properties (e.g. hasNameEN). Once again, this is done to comply with a widely used FAO convention, and to facilitate the reading of the ontologies by human users (be they ontology engineers and editors, or software developers), especially considering that typically there is more than one name for each language (usually there is a “name,” a “long name” and a “full name”). However, in this way we did not use the `rdfs:label` construction.

All pieces of information of most common use by the application for fisheries have been included in the ontologies, including the ID of all items in the database and the meta code used to identify the “type” of reference data at hand. Since the combination of ID and meta code is unique within the database we also use this combination to form the names of all instances. Moreover, meta codes and IDs are also rendered (individually) as datatype properties in order to stay aligned with current practices of interaction with the database.

All data included in the ontologies comes from the database, without further additions or modifications. As a consequence, sparsely populated columns in the database are rendered as empty properties. Since this emptiness is to be related to the dynamic nature of the database, we preferred not to impose constraints on required or non-required properties.

Some properties only make sense in conjunction with others, as in the case of geographical coordinates. In the FIGIS database, coordinates are given in order to circumscribe areas, so 4 coordinates are usually stored for each area: minimum and maximum latitude and minimum and maximum longitude. In that case it is important to distinguish each coordinate from the other, but also to group all of them together. For this reason we preferred to use one property per coordinate, and to group them as subproperties of the same superproperty.

Comments were added as much as possible to the ontologies in order to facilitate use and reuse of the ontologies.

In the next section we carefully discuss other aspects of the ontologies and of the whole network, from the point of view of its modelling and implementation.

4 Discussion

4.1.1 Ontologies of what: classification systems or real world entities?

As introduced in Sec. 3.2, a question that arose early on in our work, and that was left implicit in [D7.2.2], is whether the ontology developed should be considered ontologies of real world entities or ontologies of classification systems. Our starting point are the classification systems relevant to fisheries (which arguably do refer to real world entities), and our work mainly deals with the modelling of the fundamental concepts embodied in these classification systems (e.g. class of species as described by a taxonomic classification, where the important features to keep track of are the classification ranks and the scientific names, and the taxonomic code). However, the network also includes ontologies that seem more oriented to modelling real world entities, such as the (fragment of the) geopolitical ontology included in the network (but the same applies to the land ontology included in D7.2.2).

A different discussion should be made with ASFA reengineering work. The different semantics of RTMS elements can be singled out according to its branches, but in ASFA, we cannot establish in advance when terms refer to fisheries entities or to fisheries *types of* entities without making an extensive analysis of its terms.

Performing such an analysis is not sustainable because of the great amount of effort required, as well as because it would imply a destructive modification of the thesaurus, whose development would hardly be synchronized with the OWL version afterwards. On the contrary, we have pointed at the semantic issues underlying the reengineering patterns proposed, and have gone with the easiest and most sustainable one, i.e. the A-box reengineering pattern, which assumes a domain of interpretation constituted by the conceptual objects that have been represented by thesaurus designers as “terms”.

The recently proposed method *Semion* [Gan09] (see also Appendix VI), a method for the “reengineering on demand” of modules of ontologies (based on customizable transformation rules that explicitly declare their semantic rationale) is expected to allow the creation of domain-oriented ontology. Work will follow within WP2 to experiment with the application of such a method to ASFA (D2.4.4 [D2.4.4]).

4.1.2 Modelling the fisheries domain: how to decide between instances and classes?

Another issue that arose early on from our work on RTMS (see discussion in Sec. 3.2) is that some of the entities described in our ontologies should be better represented as classes, as opposed to using instances. A typical example is the taxonomic classification of species, as it seems obvious that real instances are only the actual animals that are caught in the see. All other entities (e.g. species, families) should better be represented as classes.

This observation is very sensible from a theoretical point of view. However, from a functional point of view, we need to be able to treat species as instances, as they are the primary object of discourse in the context of the time series collated by FIES, i.e. statistical data reported to FIES is primarily about species.

As for the higher taxa, although they are mostly used for data aggregation, it may well happen that some data is reported at a higher level than the species level (e.g. main groups) if this is convenient to the reporting body (which is for example the case when the group includes several species caught or produced in small quantities each). Moreover, given the tool at hand, the modelling of higher taxa as classes would imply that they are actually part of the ontology schema

(T-box), and therefore they would be described manually, making the process extremely error-prone, time consuming and practically non reproducible.

Moreover, learning from ASFA reengineering lessons, we observe that the domain of interpretation of fisheries can contain entities as well as types of entities, and distinguishing them in a logically-sound way would require a huge amount of fishery experts' time, and only after they are organized in a team sided by ontology designers and are taught design tools adequately. Such a task constitutes an *application of NeOn technologies*, rather than an objective of the project.

4.1.3 Beautiful modelling vs. efficient result: the case of biological entities

Biological taxonomies are a classical example of classification, also taken into consideration as an ontology design pattern in the work carried on with NeOn (see deliverable D2.5.1 [D2.5.1], Sec. 4.10.1, "Linnean Taxonomy", and Appendix A.10).

The pattern proposed by NeOn is to use two pairs of object properties to express the fact that any taxa in the taxonomy has exactly one taxa immediately above (below), and one or more higher (lower) level taxa. The object properties used would then be: "hasDirectHigherRank" ("hasDirectLowerRank") and "hasHigherRank" ("hasLowerRank"), respectively. The domain and range of these properties is the same, i.e. the most general class subsuming all the taxa. This modelling style covers the needs described in Sec. 2.3.1 and summarized in Sec. **Error! Reference source not found.**. Moreover, it applies an elegant economy of entities, since with only two pairs of object properties one can express taxonomies of any length (see Figure 14). A similar modelling could be applied to model the organization of FAO divisions of water areas (which embodies an inclusion, or part-of relation).

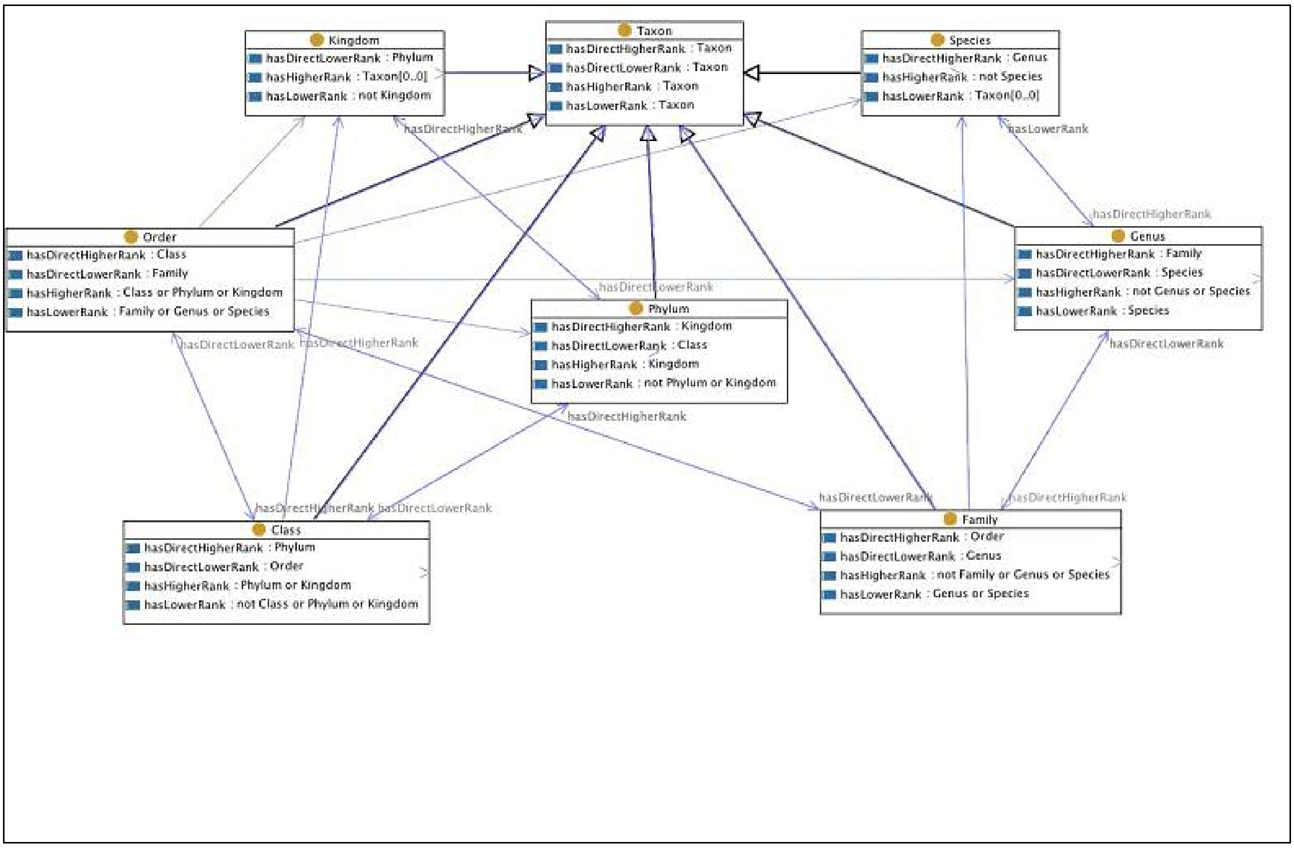

Kingdom
hasDirectLowerRank : Phylum
hasHigherRank : Taxon[0..0]
hasLowerRank : not Kingdom
Taxon
hasDirectHigherRank : Taxon
hasDirectLowerRank : Taxon
hasHigherRank : Taxon
hasLowerRank : Taxon
Species
hasDirectHigherRank : Genus
hasHigherRank : not Species
hasLowerRank : Taxon[0..0]
hasDirectHigherRank
hasLowerRank
hasDirectHigherRank
Order
hasDirectHigherRank : Class
hasDirectLowerRank : Family
hasHigherRank : Class or Phylum or Kingdom
hasLowerRank : Family or Genus or Species
Genus
hasDirectHigherRank : Family
hasDirectLowerRank : Species
hasHigherRank : not Genus or Species
hasLowerRank : Species
hasDirectLowerRank
Phylum
hasDirectHigherRank : Kingdom
hasDirectLowerRank : Class
hasHigherRank : Kingdom
hasLowerRank : not Phylum or Kingdom
hasDirectLowerRank
hasDirectHigherRank
hasDirectLowerRank
hasDirectHigherRank
hasDirectHigherRank
hasDirectLowerRank
hasDirectLowerRank
hasDirectHigherRank
Class
hasDirectHigherRank : Phylum
hasDirectLowerRank : Order
hasHigherRank : Phylum or Kingdom
hasLowerRank : not Class or Phylum or Kingdom
Family
hasDirectHigherRank : Order
hasDirectLowerRank : Genus
hasHigherRank : not Family or Genus or Species
hasLowerRank : Genus or Species

Figure 14. Design pattern for Linneus classification as described in D2.5.1.

The problem we found with implementing that model, is that the tool we used (ODEMapster) does not allow for an object property to have values belonging to more than one class (which happens with the instantiations of hasHigherRank and its inverse hasLowerRank). We hope this issue can be fully addressed in the future. Currently, we are experimenting with a work-around[58] that allowed us to extract data from RTMS according to a result *similar* model.[59] However, this work-around produced duplication of data, resulting in very large A-box files. This (partially) happens because of the "duplication" of information stemming from the fact that a direct higher rank is also a higher rank (and a direct lower rank is also a lower rank). The result is that the instance file size is 18 Mb (while the same file produced according to the less elegant model was about 13 Mb). The consequences of this are visible when dealing with it as a single ontology, as well as when it is imported to be connected with another ontology.

In the case of run-time access to the database, the observation about the size of the data file does not hold anymore. In this case, as well as in the case of the use of stand-alone the ontologies, a possible solution to the duplication issue is to skip the materialization of all inferred superproperties, leaving to reasoners the burden of materializing that information when needed. This is actually the solution applied by the Linnean taxonomy design pattern, which reuses a logical pattern called *transitive reduction*:

```
hasDirectHigherRank rdf:type owl:TransitiveProperty

hasHigherRank rdf:type owl:ObjectProperty

hasDirectHigherRank rdfs:subPropertyOf hasHigherRank
```

In this way, for each transitive closure of hasDirectHigherRank axiom, a reasoner will simply infer a hasHigherRank axiom, while we maintain the size of the A-box reasonable. For ODEMapster's sake, the single-class properties could be made subproperty of hasDirectHigherRank, so obtaining the best of the different approaches. This should be tested.

For the sake of the users of the network we also produced a second model, which embodies a less general and concise way to express the hierarchy, as it defines three properties (i.e. includesSpecies, includesFamily, includesOrder), each with only one class in its range (species, family, order, respectively).

4.1.4 Modular design driven by the amount of data (instances) available

In most cases, the ontologies (T-box) included in the network consist of very few classes and a large amount of instances. It could then be argued that some ontologies could actually be "merged" together as they do not cover distinct "domains" (and it is not possible to give a strict definition of what a domain is). For instance, this could be the case of various divisions of water bodies (FAO divisions, large marine ecosystems, exclusive economic zones), or even for the taxonomic classification of living entities and the ISSCAAP classification.

However, beyond the fact that a distinction can actually be made between these "domains" and so the modularity of our modelling can be argued for, we want to stress that one of the rationale for keeping the design modular in that way is to keep as small as possible the size of the A-box connected to the ontologies. Of course, this decision is compensated by a wider use of *owl:imports* statements.

[58] it consists in defining that the range of the property as a superclass common to all different ranges. However, this results in a duplication of entities, as explained in the following.

[59] Note that if we only used the properties hasDirectHigherRank and hasDirectLowerRank we would have only one class in the range, but we would not be able to reconstruct the entire taxonomy, since direct ranks above and below may be missing.

The modularization solution proposed in Sec. 3.3.6 is in line with the data approach, but also uses mapping and task information: whenever there is a mapping from ASFA to a modularized resource, that mapping is exploited to extract an ASFA module. Furthermore, task information (typically, expressed by competency questions) could be used with the Semion method to derive a T-box-style module that result to be optimized for that task.

4.1.5 Modularization and multiple perspectives: the case of water areas

The FAO classification of water areas is based on a classical part-of relation, strictly hierarchical and transitive. According to this classification, the entire planet is divided into areas, some of which only include water (i.e. marine waters), while others include both land and water (i.e. inland waters). However, the FAO classification does not include the information that water areas may be marine or inland.

The part-of relation embodied in the FAO classification is rendered in the RTMS by means of the usual group – member relation. Also, as one can see from Table 3, the FAO classification and the distinction of types of water (inland/marine) are encoded in different hierarchies (under the hierarchy of statistical areas, and under the hierarchy environmental area, respectively). The two hierarchies are connected by means of the group – member relation established between the items belonging to them.

We saw two possibilities to render the fact that any area may be either inland or marine. The first option is to consider “inland” and “marine” as attributes of a specific area and define a corresponding datatype property. The advantage of this modelling is that it is very compact and direct. The other option is to keep the information about the type of water apart from the actual FAO division, either in the same ontology or even in a different one. This modelling is more modular and it is compliant with the original FAO classification; it also allows for easy extensibility, as one could add more types of water, for example brackish.[60] The disadvantage is that it is slightly more laborious to create, and that the actual types of water are not part of the ontology schema, as it would be sensible to be, but extracted from the database.

Our decision was to keep the distinction between these two types of water (inland and marine) apart from the FAO classification of water bodies, but included in the very same ontology where that is described. We then defined a class “TypeOfWater”, together with an object property “isOfWaterType” with having a FAO division as domain and a type of water as range. Note however, that the observation made in the previous paragraph (about the actual types of water not being part of the ontology schema) applies.

4.1.6 When codes and data structure are not aligned: taxonomic codes

Biological items are given taxonomic codes as defined in the ASFIS list. However, these taxonomic codes do not strictly correspond to the hierarchy stored and used in the RTMS. In fact, taxonomic codes include a place for the genus, which is not actually implemented in the RTMS, as it is not used for aggregation (i.e. querying) the time series. The ontology for taxonomic classification then reflects the functional orientation of the RTMS. However, the model can be easily generalized when a clear use case is shown for it, and the corresponding data is made available to populate the ontology.

[60] One interesting outcome of this further specification is that the information about the type of water area would be available at lower level of classification than “areas”, as it currently is the case. This would open up an wide range of possibilities to enrich our knowledge base and allow finer analysis of the data available.

4.1.7 Non-hierarchical representation of hierarchical codes

Also fisheries commodities have taxonomic codes associated that embody a hierarchical classification of the items classified (e.g. ISSCFC and HS). However, the taxonomy expressed by these codes is not rendered in the RTMS hierarchy (cf. Table 5), nor encoded in the group table (which only represents the *correspondence* between each commodity item in the HS and items in the ISSCFC system). Here again, we followed the structure of the RTMS and commodities are represented in the ontology as two flat lists of instances, despite their complex taxonomic codes. However, for next release of the network we would like to investigate the possibility of rendering these taxonomic classifications also hierarchically in the ontology.

A similar observation applies to the main classification of gear types, and vessels types (ISSCFG and ISSCFV, respectively).[61]

4.1.8 Modeling multilinguality

Most of the reference data is multilingual, because it is available in more than one language, usually English, French and Spanish. However, in some cases, two or more names are available in each language, depending on where and how the name is going to be used (e.g. official short and long names for documents, short names for diagrams and captions). The best example of this is represented by the land areas, e.g. countries and organizations, as described in Sec. 2.5. In either case, names are established on the basis of international agreements that result in a 1-to-1 correspondence between languages. Official lists of these correspondences are maintained by various organizations, including FAO (cf. FAOTERM[62] and in particular NOCS[63]).

Given the need to render this extra-level of multilinguality, and to ease the “backward” connection with the source database (by staying close to existing practices adopted in FAO), we rendered object names by means of datatype properties, with the two-digit ISO code of the language added to the name of the property (as in “hasNameShortFR”). The advantage of this convention is also to ease the visualization of properties by human editors (note also the effect of this convention on alphabetic listing of the properties).

Properties are endowed with an rdf label `xml:lang` corresponding to the language at hand, and wherever possible, elements with *`rdfs:label`* are used to facilitate visualization of names in various tools.

The LIR model developed within NeOn (cf. [D2.4.1]) does cover the linguistic information covered by the RTMS ontologies from a conceptual point of view, but they differ from a structural perspective, because the linguistic/terminological information is described by different means. The difference between the structure of the linguistic part of the ontologies on the one hand, and the LIR on the other is the following: in RTMS the English label of an individual is encoded as the value of data properties such as "hasNameEN", which is associated with individuals. In LIR this label is encoded as an instance of the LIR class "Lexicalization", which is part of the class LexicalEntry. LexicalEntry is in its turn related to the individual in the ontology through the "hasLexicalization" relation.

[61] However, the case of gear types is yet different, as it is based on a filter, as opposed to a meta code. This means that ISSCFG objects do not exist as such in the item table, but can only be reconstructed by looking at the ISSCFG filter code and at the corresponding entries in the parent table. This observation is important to highlight the fact that not all data with taxonomic codes are stored in the same way, therefore special care mast be devoted to the reengineering of data sources like that into ontologies.

[62] http://www.fao.org/faoterm/index.asp?lang=EN

[63] http://www.fao.org/faoterm/nocs/pages/homeNocs.jsp?latest=1

In order to link both representations the RTMS data property needs to be converted into a LIR individual of the Lexicalization class. The detail of the conversion between them is currently under investigation at USFD.

4.1.9 Coding systems vs standard abbreviations

The ontology of taxonomic classification of biological entities features a “classification system”, the 3-alpha code, which is modelled as an attribute of the classification giving the name to the ontology. This case shows that we made a distinction between proper classification system (usually using taxonomic codes) and systems of standard abbreviations such as the 3-alpha code and the standard abbreviation used for ISSCFV types of vessels. By looking at the internal organization of RTMS we see that that latter type does not have meta codes associated.

4.1.10 A network without “previous versions”

The definition of network of ontologies included in D.1.1.5 [D1.1.5] covers the relationship of “prior version” as one of the relationship that may make a network. However, in our network the reference to prior versions (when available) is given by means of the “owl:priorVersion” statement. We did so because WP7 has no clear use case for having previous versions in the network. However, this could not be the case for next release, and clear practical indication of how to deal with such cases will have to be available.

4.1.11 Networked ontologies from theory to practice

Most interesting links between the fisheries ontologies are the instance level. As remarked in Sec. 3.4.2, our network is mainly build on the basis of linking information contained in the RTMS. The process for building the network in that case is far from straightforward, and its repetition very costly (contrary to what recommended in Sec. 1). In fact, we found that the modular style adopted in the ontology design does not correspond to the way it is implemented at the instance level, as when using ODEMapster, one has to define in the same r20 all the entities involved in the link.

The third case we identified in Sec. 3.4.2, is when linking information can be extracted directly from the ontologies (A-box). The tools available within NeOn, such as the Alignment API,[64] offer the possibility of representing alignments in a relatively open format, as in its current practical implementation, it can represent links between entities as soon as they are identified by URIs.

The representation of relations such as (species) “isFoundIn” (water) is far from standard. However, it can be used with the Alignment format with limited support (in printing and/or applying such relations). The Alignment API provides support for transforming alignments into OWL axioms (and several other languages). This, however, does only support standard OWL relations (*subClassOf*, *sameAs*). Given it current state, the Alignment API implementation can easily be further extended for better support and the team developing it is currently working on easing this process. We expect to be able to use this technology for the making of the next release of the ontology.

As for the matching of the instances, we highlight two issues:

1) most matchers are made for matching ontologies and not data. Hence, even if many of them (and in particular among NeOn partner) can provide data matching, these are not their strong points. However, there is renewed interest towards instance matching in the context of the linked data movement.

[64] http://alignapi.gforge.inria.fr/

2) the use of “non standard” relations between data is a too difficult problem for current matchers. Looking for such relations open the search space so widely that researchers prefer to avoid them. Matchers, such as the OWL Lite Alignment[65] 9OLA), are able to return instance matching, but their quality is not very high.

Some NeOn plug-ins, such as OKKAM,[66] provide support for instance naming. The investigation of their performance with FAO data is in our plan for future work. Currently, the interest in instance matching is likely to lead to a special track in the next OAEI,[67] which could bring interesting contribution to the NeOn work.

On the processing side, the point is to take advantage of alignments within the FAO context. It is possible from an alignment description to generate OWL and several other formats such as XSLT. It should be relatively easy to design new generators for the needs of FAO as soon as the relations can be expressed as an alignment.

For the inclusion of ASFA in the network, the possibilities offered by semi-automatic alignment should be exploited: because of its large size, and because of its coverage, which we can be expected to be largely overlapping with the RTMS.

4.1.12 Modular design to minimize maintenance effort?

The assumption that modularity eases the maintenance of ontologies and of the network formed by them did not prove to be strongly confirmed, especially for what concerns the A-box generation and maintenance.

In fact, during the making of the network, several versions of ontologies and data were created. However, the maintenance of the correspondences between versions of the ontologies involved in the network was not straightforward and often very time consuming and error-prone. For example, when working with ODEMapster, one has to change manually all URIs in the r20, all the import statement have to be added manually, and given the fact that all entities linked in the network have to be defined in the same r20, this brings to a large r20 file, monolithic and hard to maintain.

4.1.13 Evaluation through use

As described and discussed in Chapter 1, the ontologies here presented try to balance a number of aspects, including the requirements imposed by domain at hand, good practices for sound modelling, the possibilities provided by the tools available, “compatibility’ with the database. Also some requirements coming from the use of the RTMS in the context of the statistical time series were taken into account. However, no attempts were made at this stage to use these ontologies in a real setting, i.e. serving real applications. This investigation will start after the release of the ontologies. One application will be in FSDAS, other applications will be selected internally in FIES.

[65] http://www.iro.umontreal.ca/~owlola/alignment.html

[66] See http://www.okkam.org/test-tubes/neon-plugin. It is developed within the EC funded OKKAM project: http://www.okkam.org

[67] http://oaei.ontologymatching.org/

5 Conclusions and next steps

In this deliverable we presented (Chapter 3) and discussed (Chapter 0) the first set of fisheries ontologies created within WP7. The whole RTMS network of ontologies (T-box) is available online (http://www.fao.org/aims/aos/fi), together with the corresponding data (A-box) extracted from the database according to the model described in the ontologies. The data set was produced to test the actual possibility of accessing relational data according to ontological models and using available tools. Also, the data is made available to NeOn partners and the general public for use and reuse. The appropriate documentation is also available from the FAO same web site.

The RTMS-based ontologies have been populated with data stored in a relational database, while the links between ontologies come from the RTMS and from XML fact sheets on fisheries. The links between ontologies are expressed as object properties with domain and range described in different ontologies. The `owl:imports` mechanism is used everywhere to access these ontologies. This mechanism may become expensive, depending on the size of the imported ontologies. One possibility to minimize the impact of this point is to push forward the modularization of the network and publish separately the ontologies and the links among them. One advantage of this is also that one could more easily load ontologies in tools that request the imported ontologies to be loaded before the importing ones (as in the case of the NTK).

We succeeded in creating a network of ontologies based on the RTMS data. However, the tool available still do not allow smooth run-time access is it is, as in several occasions we had to pre-process the data, sometimes intensely. In most cases we decided to create views (which imply the use of a copy of the database, unless full rights on the database are granted). This also happens to link ontologies, when the linking data comes from different sources. Also, the current mechanism to define access to the database is rather cumbersome, hardly modular and scarcely scalable (cf. Sec. 4). Arguably, the experience we learned with the RTMS database structure compared to ontologically-oriented model, may result in useful feedback in case of future development or refinement of the data base. However, the fact that the access to the actual relational data is laborious and error prone is still an issue to be worked out. The good side of this is that the problems we came across are not theoretical but strictly related to the implementation of the tools and therefore could be easily overcome. The issue of accessing relational database by means of semantically oriented technologies is receiving increasing attention in a wide area placed at the edge of relational database technologies and semantically-oriented technologies and several tools are being developed both inside and outside NeOn. A careful analysis of and experimentation with them will certainly be part of our future work.

The ASFA reengineered ontology is based on the A-box reengineering pattern. Its schema and instances are publicly available, together with experimental mappings to SKOS and LMM, as well as sample mappings and modules (URI given in Sec. 3.3). The use of a lightweight reengineering pattern is promising for a practical synchronization with future evolution of the ASFA Thesaurus: the choices made have a clear rationale, and the ontology can still be lifted to a heavier, T-box style ontology based on adaptable transformation rules.

As for the best way of exploiting semantically-oriented technologies in the context of reference data for fisheries, with respect to both data maintenance and data exploitation, in Chapter 1 we hinted three future scenarios (ontologies only used to expose data to human users and applications, ontologies used as a means to both maintain and exploit data, data migrated to ontologies and related semantically-oriented technologies). We believe that maintenance and exploitation of the data are two fundamental issues that should receive equal attention. A deeper discussion on this, possibly on even a wider range of possibilities, will take place as soon as we gather feedback from use of the ontologies in real applications, i.e. at least the FSDAS within NeOn, and other FIES applications. This will also allow us to gather evidence on the "best" modelling styles to adopt, also with respect to the applications they are meant to serve (cf. discussion in Sec. 4.1.3, where we compared the effects of “nice” modeling vs efficient modelling --- on the basis of the available tools).

In the first network of RTMS fisheries ontologies, all ***links between ontologies are at the level of instances***. When the data about linking is in the database, we found that the creation of the network should be streamlined by improving the processor (ODEMapster) in charge of the lifting of the data. In particular: the management of the r20 instruction should be more modular; the documentation should be improved; it should be possible to have several range values per each object properties; it should be possible to automatically update URIs and add import statements; it should be possible to use `rdfs:label`.

In the case of ASFA reengineering, the sample links also start at the ASFA instance level, but do not necessarily map to RTMS instances, while several different correspondence patterns have been used (cf. Sec. 3.3.5).

Similar observations hold when the data about linking is to be extracted from other sources, such as corpora of semi-structured documents. The case in which linking between ontologies can be established from the ontologies directly, by means of links between instances, is an open and interesting field for future investigation in NeOn, and application in WP7.

This network will undergo an evaluation process similar to the one that took place after the release of the ontologies in D7.2.2. Moreover, we expect to receive useful feedback from the actual use of the ontologies (and the network, or part of it) by the NeOn partners and the public. In particular, we expect to gather feedback from the use of the network in the FSDAS and its evaluation, resulting in a dedicated deliverable (D7.7.2). The improved version of the network will be released in Deliverable D7.2.4, due at M46.

Based on work here presented, we plan on expanding the ontology on fish stock so as to accommodate data produced by an increasing number of fisheries bodies devoted to the management and protection of various fish stocks. We expect that the following ingredients are necessary: include in the network more divisions of water areas (since fisheries bodies usually adopt their own division of the water areas under their competence, or further divide existing divisions), include data about correspondences between water areas divisions, and include data about fisheries bodies.

Future work on ASFA includes including ASFA into the network by producing more links between ASFA and the RTMS-based ontologies. This could be achieved by reusing some of the mapping generated within the FOS project. Liaisons with the Aquaring project can be envisaged at this stage. It would also be interesting to experiment with alignment techniques from other partners like USFD, INRIA, and OU. The alignment made in FOS could be in this case considered as a gold standard, in order to test different automatic alignment techniques. Extensive modularization can be either performed in order to store modules that result useful for fishery tasks, or left to on-the-fly queries against the `asfad.owl` ontology. An interesting development direction is to use *application ontologies*, like FSDAS ones, in order to derive mappings, modules, and eventually T-box-level ontologies (cf. sect. 4.6) from `asfad.owl`, which can match application-oriented competency questions [D2.5.1]. Another area of improvement for ASFA is multilinguality: a possible multilingual extension can be provided after full alignment to RTMS-derived ontologies. The integration of a LIR extension (cf. D.2.4.1, T2.4) would then be highly beneficial.

We plan on experimenting more on the reuse of the geopolitical ontology, especially for what concerns groups of countries. In fact, it would be very useful to exploit the abundance of information contained in the geopolitical ontology to enrich the fisheries reference data, for example for the specification of members of regional fishery bodies (who are the main organisms monitoring and regulating catch and production of stocks)

We also plan on modelling the concept of *fishery,* which is a very challenging task. In fact, although there is no unique definition of fishery agreed upon, this concept is fundamental to all work aimed at studying, managing and making decisions and policies in the area of fisheries. Therefore, we expect that by modelling this concept and hooking it to the actual data collated by FAO (and other organizations) we would offer a very useful tool to work in the area (however, a tool does not replace or *create* neither awareness nor political will).

We will investigate the connection with geographical information systems (GIS) and the network of ontologies. In fact, many ontologies, especially those about divisions of water bodies, would profit from the integration or at least interaction with GIS data.

Finally, the technologies developed in NeOn and the lessons learned while creating the first network of fisheries ontologies can bring useful input to the area of metadata for statistical data (cf. Statistical Data and Metadata Exchange[68] (SDMX)), where a lively community is trying to achieve better data harmonization and exchange through standards.

[68] http://www.sdmx.org

Annex I. Naming conventions

URI base: www.fao.org/aims/aos/fi/

Ontology names: lower letters, words separated by underscore. The name should include the version number, in the form: "_vx_y", just before the entension.

T-box: name_ontology_vx_y.owl

A-box: name_ontology_data_vx_y.owl

Classes: capital letters, with underscore instead of spaces.

Example: http://www.fao.org/aims/aos/fi/

Properties and Relations (datatype and object properties): Camel style.

Example: hasMeta, hasCodeISO2, hasNameEN.

Instances: concatenation of "ID", meta code and ID from the database. For example: "ID_54002_105".

Annex II. Glossary of fisheries terms

Baseline The line from which the seaward limits of a State's territorial sea and certain other maritime zones of jurisdiction are measured. Normally, a sea baseline follows the low-water line (lowest astronomical tide) of a coastal State. When the coastline is deeply indented, has fringing islands or is highly unstable, *straight* baselines may be used (cf. 1982 United Nations Law of the Sea Convention (LOSC)).

Catch The total number (or weight) of fish caught by fishing operations. Commonly, one distinguishes the following types of “chatch”. For a diagrammatical representation of the concept “catch” see: ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/AnnexB1CatchConcepts.pdf

> **Product Weight** The weight of a product at the time of weighing.
>
> **Landed Weight** The mass (or weight) of a product at the time of landing, regardless of the state in which is landed (e.g. whole, or gutted or filleted). This is often the only measure available. However, it provides little indication of the mass of live fish, for this reason the landed weight is generally converted to a more meaningful measure, the most frequently used being the "Nominal catch".
>
> **Nominal Catches** The landings converted to a live weight basis by means of conversion factors. In fact it is often referred to as the "Live weight equivalent of the landings" or shortened to the "Live weight".[69]

Commodity Goods and services which are the result of production processes normally intended for sale on the market at a price that is designed to cover their costs of production.

Conversion Factor In the context of fishery statistics the term "conversion factor" is used principally when converting the volume or mass (more commonly referred to as the "weight") of a product at one stage in the production chain to its volume or mass at another stage in the chain.

Demersal Dwelling at or near the bottom of a body of water.

Economic Exclusive Zone (EEZ) The part of the marine realm seaward of territorial waters within which nations have exclusive fishing rights (cf. 1982 United Nations Law of the Sea Convention (LOSC)).

Fishery Generally, a fishery is an activity leading to harvesting of fish. It may involve capture of wild fish or raising of fish through aquaculture. A fishery can also be taken as 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 the activities.

Fishery Fleet The term "fishery fleet" or "fishery vessels" refers to mobile floating objects of any kind and size, operating in freshwater, brackish water and marine waters which are used for catching, harvesting, searching, transporting, landing, preserving and/or processing fish, shellfish and other aquatic organisms, residues and plants.

[69] In national publications the same concept is also given the name "Landings on a round, fresh basis", "Landings on a round, whole basis" or "Landings on an ex-water basis".

Fishing Vessel The term "fishing vessel" is used instead when the vessel is engaged only in catching operations.

Non-Fishing Vessel The term "non-fishing vessel" applies to vessels performing other functions related to fisheries, such as supplying, protecting, rendering assistance or conducting research or training.

Gear A fishing gear is a tool used to catch fish, such as hook and line, trawl, gill net, trap, spear.

Gross Register Tonnage (GRT) The Gross Register Tonnage represented the total measured cubic content of the permanently enclosed spaces of a vessel, with some allowances or deductions for exempt spaces such as living quarters (1 gross register ton = 100 cubic feet = 2.83 cubic metres).

Gross Tonnage (GT) The Gross Tonnage for ships of 24 metres in length and over refers to the volume of all ship's enclosed spaces (from keel to funnel) measured to the outside of the hull framing.

Inland Waters The surface water existing inland, including lakes, ponds, streams, rivers, natural or artificial watercourses and reservoirs, coastal lagoons, artificial water bodies and other land-locked (usually freshwater) waters (such as the Caspian See, Aral Sea, ...). (Cf. Marine waters, Internal waters)

Internal waters Those waters of the sea on the landward side of the baseline used by the national authorities of the coastal country to measure further seawards the width of the territorial sea and any adjacent marine waters, whether salt, brackish, or fresh in character. Such "internal" marine waters will be found, for instance, when the baselines are drawn across the mouths of bays or along a "curtain" of islands lying close off the coast. Japan's well-known "Inland Sea" is not part of that country's inland waters but is one of the internal waters of Japan and forms part of the truly marine fishing areas of that country. [Article 8 of the Informal Composite Negotiating Text / Revision 2 (A/CONF.62/WP.10/Rev. 2, 11 April 1980) of the United Nations Third Conference on the Law of the Sea - This UNCLOS]

Large marine ecosystems (LME) Region (the order of 200,000 km² or greater) of the world's ocean, encompassing coastal areas from river basins and estuaries to the seaward boundaries of continental shelves and the outer margins of the major oceans current systems. The system of LMEs has been developed by the US National Oceanic and Atmospheric Administration (NOAA) to identify areas of the oceans for conservation purposes.

Marine Waters Oceans and seas including adjacent saltwater areas.

Nominal Catch The sum of the catches that are landed (expressed as live weight equivalent). Nominal catches do not include unreported discards.

Production The total living matter (biomass) produced by a stock through growth and recruitment in a given unit of time (e.g. daily, annual production). The "net production" is the net amount of living matter added to the stock during the time period, after deduction of biomass losses through mortality. Also: The total elaboration of new body substance in a stock in a unit of time, irrespective of whether or not it survives to the end of that time. Also called: net production.

Stock The exploitable group of individuals of the same species existing in a particular area at a particular time.

Straddling stocks Straddling stocks are stocks of fish such as pollock, which migrate between, or occur in both, the economic exclusive zone (EEZ) of one or more States and the high seas. Thus, the definition also includes highly migratory fish stocks.

Territorial waters The belt of coastal waters extending at most 12 nautical miles from the baseline of a coastal state. (cf. 1982 United Nations Law of the Sea Convention (LOSC))

Time units The normal annual time unit used in fishery statistics is the calendar (or civil) year: between 1 January and 31 December. However, for certain specific purposes (e.g. for Antarctic pelagic whaling fisheries; and fiscal purposes) it is deemed more appropriate to use a split year. For Antarctic pelagic whaling fisheries, the split year is 1 July-30 June. Countries using a split year (1 July – 30 June) are: Australia, Bangladesh, Myanmar, Nepal, US Virgin Islands. The Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) covering the FAO Major Fishing Areas 48, 58 and 88, used to collect data on a split year basis (1 July - 30 June) up to June 2002. After that, the "CCAMLR fishing season" was adopted for reporting fishery activities (1 December - 30 November). In tabulations where space restricts the labelling of a split year to a single year or where data for calendar and split years are tabulated together, the practice is for the split year to be represented by the calendar year in which the split year ends.

Annex III. List of acronyms

ASFIS Aquatic Sciences and Fisheries Information System

ASFA Aquatic Science and Fisheries Abstracts

CWP Coordinating Working Party on Fishery Statistics

EEZ Exclusive Economic Zone

FIES FAO Fisheries and Aquatic Information and Statistical Service

FIRMS Fishery Resources Monitoring System

GAUL Global Administrative Unit Layer

GRT Gross Registered Tonnage

GT Gross Tonnage

HS Harmonized Commodity Description and Coding System

ISO International Organization for Standardization

ISSCAAP International Standard Statistical Classification of Aquatic Animals and Plants

ISSCFC International Standard Statistical Classification of Fishery Commodities

ISSCFG International Standard Statistical Classification of Fishing Gears

ISSCFV International Standard Statistical Classification of Fishing Vessels

KOS Knowledge Organization Systems

LME Large Marine Ecosystems

LMM Linguistic Metamodel

NOAA US National Oceanic and Atmospheric Administration

RT Reference Tables

RTMS Reference Tables Management System

SITC Standard International Trade Classification of the UN

SKOS Simple Knowledge Organization Systems

Annex IV. Correspondences between RTMS dabatase and ontologies

All talbes in RTMS share the same table with the hierarchy of meta codes: md_refobject,

Taxonomic classification

Item table: fic_item. meta=31001,

Group table: fic_item_grp (currently this is a vew)

Exclusive economic zones (EEZ)

ref_water_area

Gears

ref_gear_type, ref_gear_type_grp

ISSCAAP classification

Item table: fic_item

Group table: fic_item_grp (currently this is a vew)

ISSCFC_HS commodities

items: ref_commodity_item.

Correspondinces between classifications: ref_commodity_harmonized, ref_commodity_issacaap (view), ref_commodity_fao, ref_commodity_grp.

Large marine ecosystems

ref_water_area

Vessels.

ref_vesselclass, ref_vessel_type

Water FAO divisions

Item table: fic_catch_area

Group: fic_catch_area_agg_grp

Connections between ontologies:

ISSCAAP_classification and taxonomic_classification

fic_item_grp

ISSCAAP_classification and ISSCFC_HS

ref_commodity_grp

large marine ecosystems (LME) and FAO fishing areas.

water_area_intersection

Country Exclusive Economic (EEZ) and FAO fishing areas.

water_area_intersection

Annex V. Hierarchy of types in the DB

This Annex contains the fragment of the hierarchy of types used in RTMS, represented as indented tree. The high level hierarchy is represented in Table 2 (cf. Sec. 3.1).

- 1 figis object
    - 20 000 Water area
        - 21 000 Environmental area
            - 21 001 Inland/Marine
            - 21002 Ocean
            - 21003 North/South Equatorial
            - 21004 Sub Ocean
            - 21005 Large Marine Ecosystem
        - 22 000 Fishing Statistical area
            - 22 001 FAO statistical area
                - 22 010 FAO major fishing area
                - 22 020 Subarea
                - 22 030 Division
                - 22 040 Subdivision
                - 22 050 Subunit
            - 23 000 Areal grid system
                - 23 002 5 degree square
        - 24 020 Jurisdiction area
            - 24 023 Country exclusive economic zone (EEZ)
- 29 999 Area filter

Table 3. Fragment of the FIGIS tree concerning the water areas.

- 1 figis object
  - 30 000 Biological entity
    - 31 000 Taxonomic entity
      - **31 001** Group
      - **31 002** Order
      - **31 003** Family
      - **31 005** Species
        - 31 105 Species with factsheets
    - 32 000 Commercial group of species
      - 32 001 ISSCAAP group
      - 32 002 ISSCAAP division

Table 4. Fragment of the FIGIS hierarchy concerning biological entities.

- 1 figis object
  - 45 000 Fishery commodity
    - 45 100 Commodity item
    - 45 200 Commodity harmonized group
    - 45 300 Commodity ISSCAAP group/division
    - 45 400 Commodity FAO group

Table 5. Fragment of the hierarchy of meta relative to fisheries commodities.

- 1 figis object
  - 50 000 Gear type
    - 51 000 International
      - 54 000 Gear category
      - 54 001 Gear subcategory
    - 54005 ISSCFG
  - 60 000 Vessel size categories
    - 61 000 Vessel length classification
      - 61 010 Vessel length class
    - 62 000 Vessel GRT classification
      - 62 020 Vessel GRT division
  - 64 000 Vessel type
    - 64 200 Vessel category

Table 6. A fragment of the hierarchy of meta codes (those used are in bold).

- 1 figis object
    - 190 000 Resource
        - 195 000 Resource object

Table 7. The fragment of the figis hierarchy relative to the stocks, or fisheries resources.

Annex VI. The reengineering of ASFA

Examples from asfad.owl

We include here some examples from `asfad.owl`, which contains ASFA terms and their relations (see Figure 15):

```
(25) asfad:Tidal_barrages rdf:type asfam:Descriptor
(25) asfad:Tidal_barrages asfam:BT asfad:Barrages
(26) asfad:Tidal_barrages asfam:RT asfad:TidalPower
(25) asfad:Farm_ponds rdf:type asfam:NonDescriptor
(27) asfad:Farm_ponds asfam:use asfad:Fish_ponds
(28) asfad_Undercurrents asfam:NT asfad:Water_currents
```

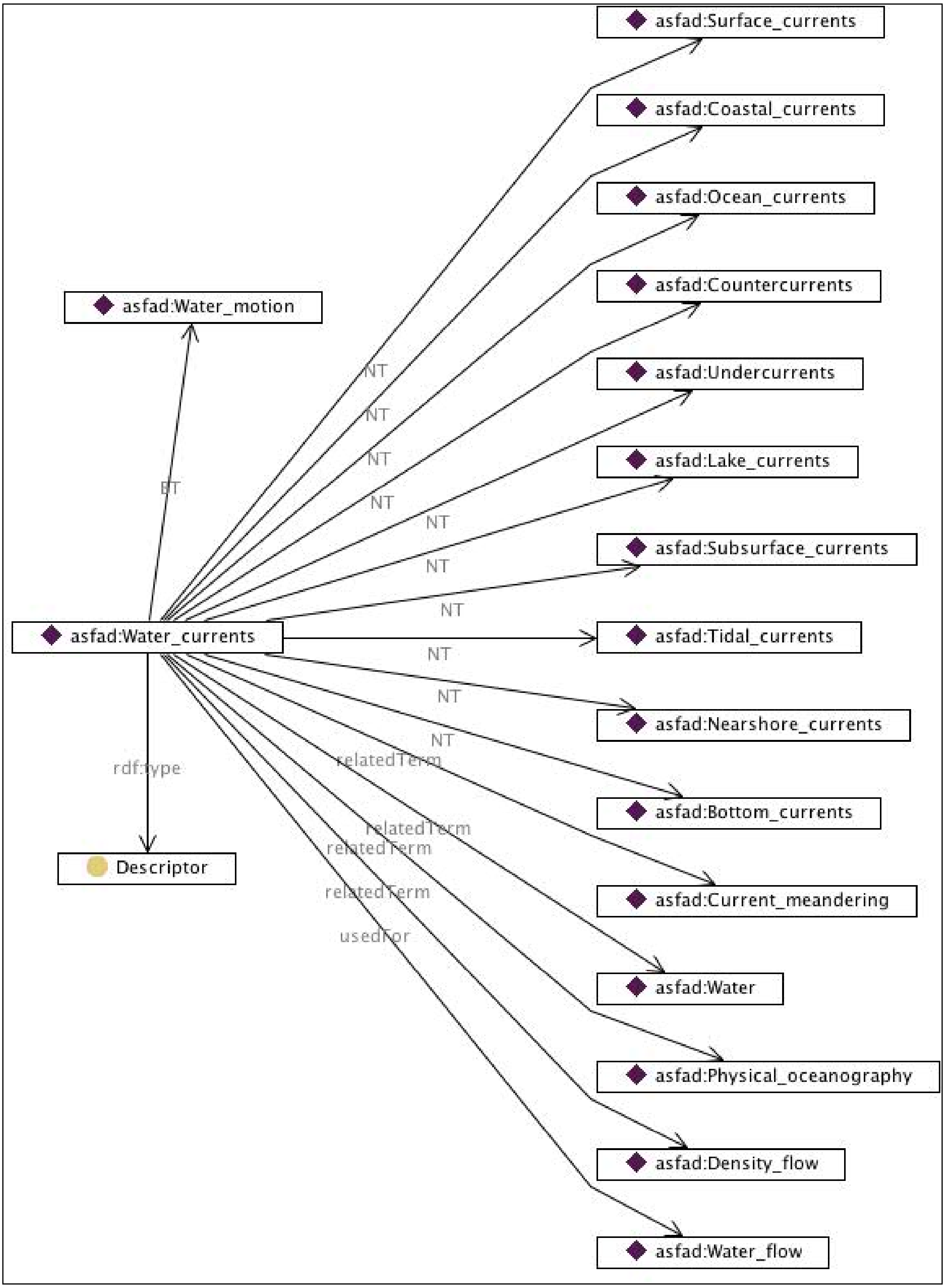


Figure 15. An example of a descriptor from asfad.owl with BT, NT, and RT relations.

An additional example by using *Semion* method [D2.4.4] is the following. Firstly, we create a new namespace for the elements of the new ontology, e.g.:

```
http://www.ontologydesignpatterns.org/ont/fao/asfa/tidalbarrages.owl
```

and import `asfam2lmm.owl` to it:

```
tidalbarrages.owl owl:imports asfam2lmm.owl
```

Secondly, we define custom transformation rules (via SPARQL CONSTRUCT) from LMM elements to OWL elements, with a new name for an `owl:ObjectProperty` that can act as a generic property for `asfam:relatedTerm` axioms.

Rule 1: create an `owl:Class` for each element that is both a `lmm1:Meaning` and a `skos:Concept` after running an A-box reasoner (e.g. Pellet, OWLIM):

```
CONSTRUCT {

?x rdf:type owl:Class .
}
Where {
?x rdf:type lmm1:Meaning .
?x rdf:type skos:Concept .
}
```

Rule 2: create a `rdfs:subClassOf` axiom for each element that is both a `lmm1:Meaning` and a `skos:Concept`, and `lmm1:specializes` another element after running an A-box reasoner:

```
CONSTRUCT {

?x rdfs:subClassOf ?y .
}
WHERE {
?x rdf:type lmm1:Meaning .
?x rdf:type skos:Concept .
?x lmm1:specializes ?y .
}
```

Rule 3: create an `owl:Restriction`, and a `rdfs:subClassOf` axiom to that restriction, for each element that is both a `lmm1:Meaning` and a `skos:Concept`, and has a `asfam:relatedTerm` axiom after running an A-box reasoner:

```
CONSTRUCT {

?x rdfs:subClassOf _:a .
_:a rdf:type owl:Restriction  .
_:a owl:onProperty lmm1:relatedMeaning .
_:a owl:someValuesFrom ?y
}
WHERE {
?x rdf:type lmm1:Meaning .
?x rdf:type skos:Concept .
?x asfam:relatedTerm ?y
}
```

After firing the rules and asserting the results into `tidalbarrages.owl`, the result of this instantiation of *Semion* is the following set of axioms:

```
(29) tidalbarrages:Tidal_barrages rdf:type owl:Class

(30) tidalbarrages:Tidal_barrages rdfs:subClassOf tidalbarrages:Barrages

(31) tidalbarrages:Tidal_barrages rdfs:subClassOf

     (lmm1:relatedMeaning some tidalbarrages:TidalPower)
```

The alignment of SKOS to the linguistic metamodel (LMM)

SKOS has been independently aligned to LMM (Linguistic MetaModel) [PGG][Gan09], which is also aligned to many other resources (Wordnet, FrameNet, LMF, DBpedia, etc.), as well as to the `codolight.owl` ontology design metamodel [D2.1.2], used in order to integrate the different models and tool descriptions developed in NeOn. The advantages of aligning SKOS (and then ASFA via inheritance) to LMM include:

- the smooth navigation between heterogeneous KOSes, lexica, and relevant knowledge resources across the Semantic Web;
- the custom reengineering of modules extracted from `asfad.owl` into regular T-box domain ontologies, by applying the *Semion* method.

Semion has been theoretically described in [Gan09], and its implementation with respect to NeOn reengineering methods will be fully introduced in deliverable D2.4.4 [D2.4.4]. The basic LMM vocabulary has the following namespace:

`http://www.ontologydesignpatterns.org/ont/lmm/LMM_L1.owl`

The SKOS alignment to LMM contains the following axioms, which hold also for the `asfam.owl` classes and properties that are `rdfs:subClassOf` or `rdfs:subPropertyOf` respectively of SKOS ones (Figure 16):

- `skos:Concept rdfs:subClassOf lmm1:Meaning`
- `skos:semanticRelation`[70] `rdfs:subPropertyOf lmm1:relatedMeaning`
- `skos:broader rdfs:subPropertyOf lmm1:specializes`
- `skos:narrower rdfs:subPropertyOf lmm1:isSpecializedBy`
- `skosmapping:exactMatch rdfs:subPropertyOf lmm1:relatedMeaning`

[70] `skos:broader, skos:narrower`, and `skos: related` are `rdfs:subpropertyOf skos:semanticRelation`

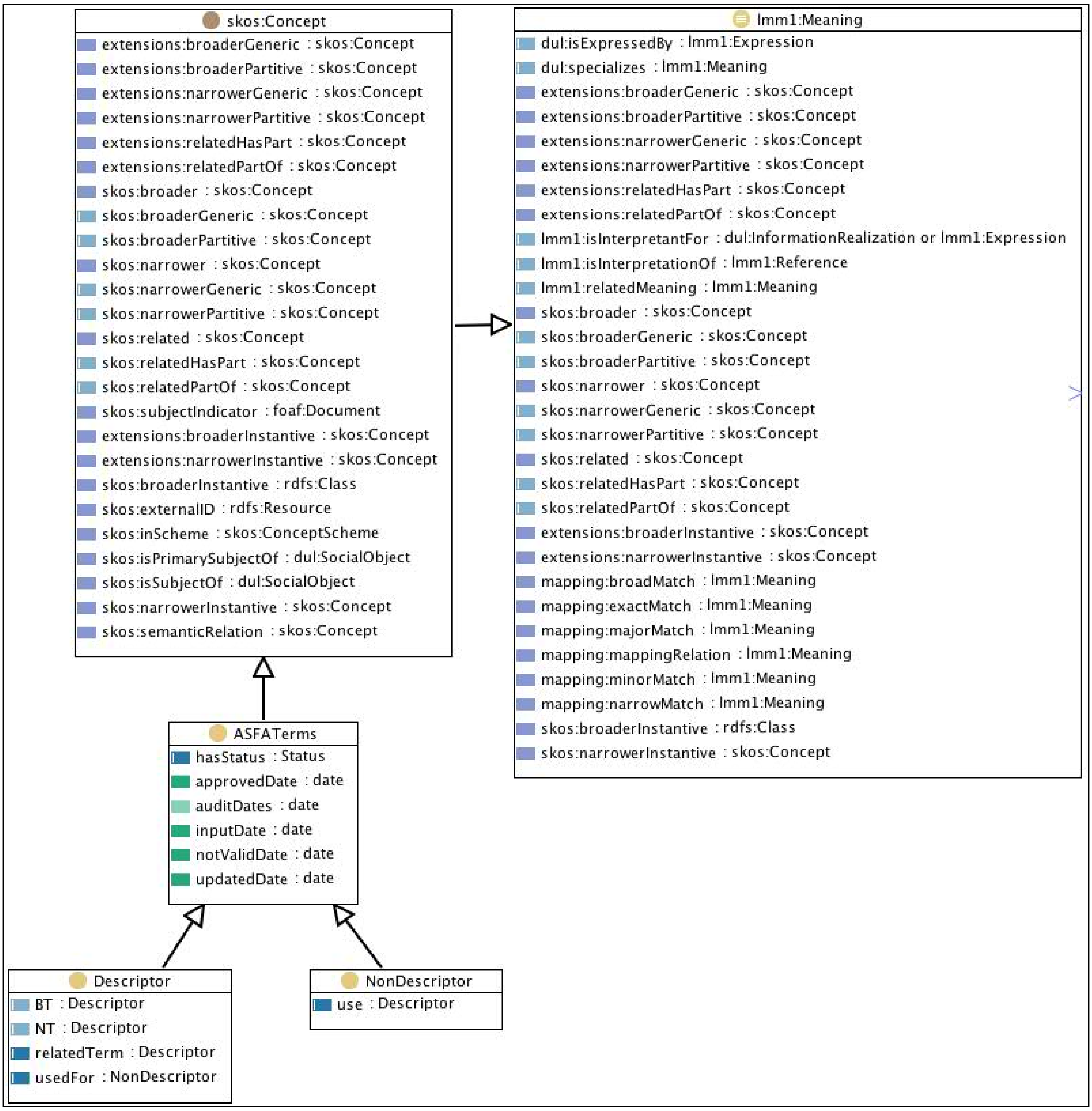

skos:Concept
extensions:broaderGeneric : skos:Concept
extensions:broaderPartitive : skos:Concept
extensions:narrowerGeneric : skos:Concept
extensions:narrowerPartitive : skos:Concept
extensions:relatedHasPart : skos:Concept
extensions:relatedPartOf : skos:Concept
skos:broader : skos:Concept
skos:broaderGeneric : skos:Concept
skos:broaderPartitive : skos:Concept
skos:narrower : skos:Concept
skos:narrowerGeneric : skos:Concept
skos:narrowerPartitive : skos:Concept
skos:related : skos:Concept
skos:relatedHasPart : skos:Concept
skos:relatedPartOf : skos:Concept
skos:subjectIndicator : foaf:Document
extensions:broaderInstantive : skos:Concept
extensions:narrowerInstantive : skos:Concept
skos:broaderInstantive : rdfs:Class
skos:externalID : rdfs:Resource
skos:inScheme : skos:ConceptScheme
skos:isPrimarySubjectOf : dul:SocialObject
skos:isSubjectOf : dul:SocialObject
skos:narrowerInstantive : skos:Concept
skos:semanticRelation : skos:Concept
lmm1:Meaning
dul:isExpressedBy : lmm1:Expression
dul:specializes : lmm1:Meaning
extensions:broaderGeneric : skos:Concept
extensions:broaderPartitive : skos:Concept
extensions:narrowerGeneric : skos:Concept
extensions:narrowerPartitive : skos:Concept
extensions:relatedHasPart : skos:Concept
extensions:relatedPartOf : skos:Concept
lmm1:isInterpretantFor : dul:InformationRealization or lmm1:Expression
lmm1:isInterpretationOf : lmm1:Reference
lmm1:relatedMeaning : lmm1:Meaning
skos:broader : skos:Concept
skos:broaderGeneric : skos:Concept
skos:broaderPartitive : skos:Concept
skos:narrower : skos:Concept
skos:narrowerGeneric : skos:Concept
skos:narrowerPartitive : skos:Concept
skos:related : skos:Concept
skos:relatedHasPart : skos:Concept
skos:relatedPartOf : skos:Concept
extensions:broaderInstantive : skos:Concept
extensions:narrowerInstantive : skos:Concept
mapping:broadMatch : lmm1:Meaning
mapping:exactMatch : lmm1:Meaning
mapping:majorMatch : lmm1:Meaning
mapping:mappingRelation : lmm1:Meaning
mapping:minorMatch : lmm1:Meaning
mapping:narrowMatch : lmm1:Meaning
skos:broaderInstantive : rdfs:Class
skos:narrowerInstantive : skos:Concept
ASFATerms
hasStatus : Status
approvedDate : date
auditDates : date
inputDate : date
notValidDate : date
updatedDate : date
Descriptor
BT : Descriptor
NT : Descriptor
relatedTerm : Descriptor
usedFor : NonDescriptor
NonDescriptor
use : Descriptor

Figure 16. The ASFATerms class and subclasses in asfam.owl as aligned to LMM.

Annex VII. Addenda based on Y3 review

This Annex has been added in order to address the request made by the EU reviewers. The request is the following:

"The deliverable should provide a summary on what NeOn plugins and methodology are used in the process of constructing the network of ontologies, as well as which features of the Toolkit are necessary but not offered by other ontology editors and which features are the most needed ones for further development of the network of ontologies. Does NeOn actually provide a better environment than competing offerings for the challenges described in this report?"

We address these questions in a separate annex in order to facilitate reviewers in reading the answers provided. Necessary cross-reference have been added between this annex and the rest of the document, and vice-versa.

Plugins and methodology used in the process of constructing the network of ontologies

We based our work on the NeOn methodologies described in NeOn deliverables D5.4.1 and D5.4.2, focussing on the reengineering non-ontological resources (cf. scenario 2 described in D5.4.2, Chpt. 3). The preliminary phase of ontology specification[71] was already carried out for at the time of the first release of fisheries ontologies (cf. [D7.1.1, D7.1.2]). In particular, we identified: 1) the purpose, 2) the intended uses and users of the ontologies, and 3) the set of ontology requirements that the ontology should satisfy after being implemented. The NeOn fisheries ontologies based on reference data are envisaged to continue serve as reference data for statistics, and to be used by other applications able to exploit the semantics contained in them, primarily the FSDAS. The intended human users are ontology developers and domain experts (in marine biology, environment, oceanography, economy, fisheries) who take care of the data maintenance. Besides the human users we also contemplate non-human users, namely the applications that are going to exploit the data. Easy maintenance by humans and exploitation by application are the primary requirements for the ontologies after implementation.

The NeOn methodology relative to the reengineering of non-ontological resources consists in the following activities: 1) non-ontological resource reverse engineering, 2) non-ontological resource transformation, 3) ontology forward engineering. We applied this schema in a highly iterative way, by repeating the cycle of phase 1-2-3 and adding some mini-evaluation, usually rather informal, to check how close we were to the wished result. Phase 1) required extensive work in order to gain the proper understanding of the domain to be modelled and of the data to be reengineered. Conversations with statistical officers and experts of metadata for statistics and international coding systems helped gain the needed understanding of the statistical data gathered by FAO, of the way it is collected (statistical data is provided by member countries and other international

[71] ***Ontology Specification*** is defined in D5.3.1 as a collection of requirements that the ontology should meet. The output of this activity is the ontology requirements specification document (ORSD) that includes the purpose, level of formality and scope of the ontology, target group and intended uses of the ontology, and a set of requirements, which are those needs that the ontology to be built should cover.

bodies) and of its metadata. Repeated conversations with domain experts were fundamental to gain understanding of the specific subdomains at hand: biologists and oceanographers for the taxonomies of biological entities; fisheries experts for the classifications of vessels and gear types; economists for the construction and use of fisheries commodities; GIS, legal officers and managers for the divisions of water bodies. The “summary” sections at the end of each subsection in Sec. 2 (in this deliverable) result from the work carried out in phase 1 -- and in view of phase 2. The lessons learned discussed in Sec. 4 result from the work carried out in phase 2 and 3. An evaluation phase follows the release of the deliverable, and will be included in next deliverable, D7.2.4, due at M47.

The population of the ontologies in the network was achieved by using ODEMapster, a tool that is now integrated as a plugin for the NTK. However, we mostly used ODEMapster as a stand-alone tool accessed from command line, because at the time of the work, the integrated plugin did not provide access to all functionalities of the tool. ODEMapster was also used to populate networked ontologies, in all cases when linking information is in the DB (cf. discussion in Sec. 3.4.2). When linking information was not directly available from the DB (e.g. contained in fact sheets), a more complex procedure had to followed, consisting of: a step of information extraction from the relevant source, a step of data manipulation to make it compatible with the data sets to be linked (typically based on converting between classification systems), a step of format harmonization. Since processes of that type are not supported by NTK, we used a combination of tools in order to produce those parts of the network: the first step was achieved by using XLST technologies applied to the XML fact sheets; the second step involved ad-hoc queries to the DB, the third step was often based on ODEMapster, after storing the relevant information in a relational table. At the time of our work, alignment plugins were still not satisfactorily integrated in the NTK, so we did not experiment with them.

As for the visualization of the network, we make extensive use of HTML visualization to show ontologies to domain experts (a big advantage of this visualization, is that since it only requires a browser, users are not forced to install and learn any new tools). This functionality is provided by the OWLDoc plugin for NTK, which is preferable to the competitor OWLDoc for Protégé because of many features (e.g. visualization of class and property trees, and visualization of individual pieces of information concerning instances), but it still behind in giving an overall view of the network (the OWLDoc for NTK shows all classes together, while OWLDoc for Protégé shows classes in separate boxes, one per ontology). For this reasons, every time the network information were more important than information about individual ontologies, we used the OWLDoc plugin for Protégé. We also felt the need to produce an overall visualization of the entire network (cf. Figure 13), which we produced manually, as it was not available from within the NTK (to the best of our knowledge, no other tools offer this type of visualization). This feature is going to be introduced in the next release of the NTK.

Needed features for further development of network of ontologies

The key needed features to allow for further development of networks of ontologies (based on non-ontological resources) are the following:

1. efficient, reliable access and/or export of data (e.g. relational data) according to the ontological framework;
2. efficient data maintenance, with support to various user profiles and roles in the process;
3. support to ontology networking available to domain experts;
4. reliable versioning policy and adequate support for data linking.

Access and reengineering of non-ontological resources (RDB). The issue of connecting to the database, specifically to the relational database used in our case study, has been discussed at length in this deliverable. Our conclusions is that the process of accessing data from the DB is not quite straightforward, but the difficulties are of only technical nature: some of the problems encountered may be addressed with appropriate simplification of the DB, others by further developing the functionalities already provided by the NTK, others by improving the usability of the toolkit and of the relevant plugin (ODEMapster) and providing appropriate documentation.

Data maintenance. The phase of data maintenance (usually performed by domain experts) usually represents the biggest part of any data set lifecycle. Therefore, the quality of the maintenance support often decides for the adoption or rejection of a system in a real-life environment. Here, by “data” we mean the ontological data (networked ontologies, currently expressed in OWL), their instances (currently, expressed in RDF) and links between data. In scenarios like ours (international organizations, with several roles contributing to different aspects of the same project) the former activity is typically performed by ontology experts and is currently supported by most ontology editors, while the latter activities are performed by domain experts (i.e. non-ontology experts), according to a specific editorial workflow. When maintaining instances, then, it is important that an appropriate editorial workflow is supported, the information is visualized in a domain expert-friendly manner, and that the entire human-computer interaction model is also appropriate. Similar considerations apply when considering the linking information between instances. In the work presented in this deliverable, most linking information was extracted from the relational database, therefore it creation and maintenance was done by ontology experts (as opposed to domain experts). In order for this task to be performed by domain experts, it is necessary that data visualization is tailored for these users, and that facilities for information extractions from different sources are available to them (cf. discussion in Sec. 3.4.2).

Besides these concerns, the issue of data storage also deserve some words. Currently, when data is stored in a RDBMS, the ontology network based on that is an ad-hoc production from it, in the sense that if data changes in the database, ontologies need to be re-populated[72]. This means that some phases of the data lifecycle are not part of the ontology lifecycle. In order for the data lifecycle to coincide with the ontology lifecycle, the following scenarios may be envisaged: in the first scenario, maintenance is kept in the DB and ontologies are used only as an interface for editors to manipulate the DB (applications use ontologies to access data in a richer form than relational form). This option is not part of future development of the NTK. The second option is that, after the necessary tests, data is migrated in an ontological format (RDF/OWL) and maintained as part of an ontology network. In order for this scenario to be realized, the ontology editor must be very efficient and reliable concerning to the data storage and manipulation, and very user-friendly. Moreover, convincing evidence should be gained concerning the data exploitation phase.

Versioning. The issue of versioning is felt as crucial. In this work we concentrated on reengineering non-ontological data into networked ontologies, but we realized that strong policies of data versioning need to be in place, especially concerning linking information, and possibly the network as a whole. A more extensive discussion on this issue is going to be part of next deliverable, D7.2.4.

NTK-specific features needed to construct a network of ontologies not offered by others

The NTK is currently ahead of competitors for these features:

1. navigation of several ontologies networked together (although speed is sometimes an issue);
2. availability of ontology design patterns;

[72] Here we do not distinguish between batch and run-time access to the database, as our focus is on data maintenance.

3. connecting to non-ontological sources, namely relational databases (although still to be synchronized with ontology design patterns);
4. support of workflow editing (at the time of writing, in a pre-release phase);
5. support for ontology alignment.

Does NTK provide a better environment than competitors for the challenges described in this report?

The task of supporting the entire lifecycle of a network of ontologies (i.e. creation, population, maintenance, dissemination, exploitation of networked ontologies) is currently not entirely achieved by the NTK. However, the following considerations are due: on the one hand, the NTK supports a larger portion of these tasks than competitors, on the other hand, currently NTK presents some problems that may induce users to go for competitors who may offer “less” (fewer functionalities, or less powerful) but in a more usable, flexible and friendly manner. Our conclusion is that NTK may become the preferred choice over competitors, but still has not proven to be mature enough to be so for use in a real setting.

As we discussed in previous sections, NTK offers a number of functionalities not currently offered by competitors. However, the overall interaction with the system is somewhat cumbersome, which may prevent users to feel comfortable with the system. Compared to other ontology editors, NTK suffers from the following limitations: it is rather heavy, rather rigid in its user interaction, little tolerant to errors in the files, not very informative when problems are encountered, and it tends to be slow with big data sets. Access to non-ontological sources need to be improved and smoothly integrated with ontology design patterns. Functionalities for ontology alignment are promising, but, at the time of our work, still to be smoothly integrated with other activities of data maintenance[73], also, according to our experience, functionalities for linking instances are missing (cf. Sec. 3.2.4).

[73] Given that our experience in this respect is limited, we prefer to postpone a deeper discussion on this topic to next deliverable, D7.2.4.

References

[AGMES] FAO. Agricultural Metadata Element Set. http://www.fao.org/aims/agmes_intro.jsp

[AIDA] http://www.fao.org/fi/figis/devcon/schema/3_6/aida.xsd

[AGROVOC] FAO. AGROVOC thesaurus. http://www.fao.org/aims/ag_intro.htm

[ASFA] FAO. ASFA thesaurus. http://www4.fao.org/asfa/asfa.htm

[ASFIS] L. Garibaldi, S. Busilacchi. ASFIS List of species for fishery statistic purposes. ftp://ftp.fao.org/docrep/fao/006/y7527t/y7527t00.pdf

[BAR03] J. Barrasa and O. Corcho and A. Gomez-Perez. Fundfinder -- a case study of database-to-ontology mapping. In Proc. ISWC Semantic integration workshop. 2003.

[BAR06] J. Barrasa. Semantic upgrade and publication of legacy data, Ontologies for Software Engineering and Software Technology, 2006

[BAR07] J. Barrasa. Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales. PhD thesis. Facultad de Informativa, Universidad Politecnica de Madrid. Madrid, Spain. March 2007. Isbn: 90-75176-81-3. http://eprints.eemcs.utwente.nl/7146/

[BIZ03] C. Bizer. D2R MAP - A Database to RDF Mapping Language. In Proc. of 12th International World Wide Web Conference. 2003.

[DC] Dublin Core Metadata Initiative. http://dublincore.org/

[DCT] Dublin Core Terms. http://dublincore.org/documents/dcmi-terms/

[D1.1.1] D1.1.1. Networked Ontology Model. NeOn project report. http://www.neon-project.org/web-content/index.php?option=com_weblinks&catid=17&Itemid=35

[D7.1.1] D7.1.1. Specification of user requirements on the case study. 2006. NeOn project report. http://www.neon-project.org/web-content/index.php?option=com_weblinks&catid=17&Itemid=35

[D1.1.5] D1.1.5. Updated version of the network ontology model. Forthcoming.

[D2.1.2] D2.1.2. The collaborative ontology design ontology (v2). 2009.

[D2.2.1] D2.2.1: Methods for Selection and Integration of Reusable Components from Formal or Informal User Specifications. NeOn Project Deliverable, available at http://www.neon-project.org (2007).

[D2.4.1] D2.4.1. Multilingual ontology support. http://www.neon-project.org/web-content/index.php?option=com_weblinks&view=category&id=17&Itemid=73

[D2.4.4] D2.4.4: An integrated model for lexical/terminological resources and ontologies. NeOn Project Deliverable, forthcoming (2009).

[D2.5.1] D2.5.1. A Library of Ontology Design Patterns. http://www.neon-project.org/web-content/images/Publications/neon_2008_d2.5.1.pdf

[D5.4.1] D5.4.1. NeOn methodology for building contextualized ontology networks. http://www.neon-project.org/web-content/images/Publications/neon_2008_d5.4.1.pdf

[D5.4.2] D5.4.2. Revision and extension of the NeOn methodology for building contextualized ontology networks. http://www.neon-project.org/web-content/images/Publications/neon_2009_d542.pdf

[D7.2.1] D7.2.1. Inventory of fishery resources and information management systems. 2007. NeOn project report. http://www.neon-project.org/web-content/index.php?option=com_weblinks&catid=17&Itemid=35

[D2.1.1] D2.1.1. Design rationale for collaborative development of networked ontologies. NeOn deliverable. February 2007.

[D2.5.1] D2.5.1. Library of formal models and design patterns for collaborative development of networked ontologies. NeOn project report.

[D7.4.1] D7.4.1. Software Architecture for managing the Fisheries Ontologies Lifecycle. NeOn project report. To appear.

[D7.2.2] D7.2.2. Revised/enhanced Fisheries ontologies. http://www.neon-project.org/web-content/images/Publications/neon_2007_d7.2.2.pdf

[D7.6.2] D7.6.2. Second prototype of the FSDAS. Forthcoming.

[EDC] Extended Dublin Core. http://dublincore.org/schemas/xmls/qdc/2003/04/02/dc.xsd

[FA] FAO. Fisheries Fact Sheet. http://www.fao.org/fi/website/FISearch.do?dom=factsheets

[FAOdiv] CWP Handbook of Fishery Statistical Standards. Fishing Areas for Statistical Purposes. http://www.fao.org/fi/website/FIRetrieveAction.do?dom=ontology&xml=sectionH.xml

[FISTAT] FAO. Fisheries and Aquaculture Department. Statistics. http://www.fao.org/fi/website/FIRetrieveAction.do?dom=topic&fid=16062

[FS] FAO. Fisheries fact sheets. http://www.fao.org/fi/website/FIRetrieveAction.do?dom=topic&fid=16062&lang=en

[FSdic] FIGIS XML. List of elements. http://www.fao.org/fi/figis/devcon/diXionary/figisdoc3.5.html

[FSschema] XML schema for Fisheries Fact Sheets. http://www.fao.org/fi/figis/devcon/schema/3_6/fi.xsd

[GAN04WW] Gangemi A. WonderWeb Deliverable D16: “Reusing semi-structured terminologies for ontology building: A realistic case study in fishery information systems“, http://wonderweb.semanticweb.org, 2004.

[Gan09] A. Gangemi. What's in a Schema. In C.R. Huang and N. Calzolari and A. Gangemi and A. Lenci and A. Oltramari and L. Prevot (eds.): Ontologies and the Lexicon, Cambridge University Press, Cambridge, UK, 2009.

[GFK+04] A. Gangemi, F. Fisseha, J. Keizer, I. Pettman, and M. Taconet. A Core Ontology of Fishery and its use in the Fishery Ontology Ser vice Project. In First International Workshop on Core Ontologies, EKAW Conference, CEUR-WS, volume 118, 2004.

[HBFSS] Coordinating Working Party on Fishery Statistics (CWP). CWP Handbook of Fishery Statistical Standards. Partially available at: http://www.fao.org/fi/website/FISearch.do?dom=ontology

[HDL] G. De Giacomo, M. Lenzerini, R. Rosati. Towards Higher-Order DL-Lite. Proc. of DL2008, 2008.

[HS07] World Customs Organizations. Harmonized Commodity Description and Coding System. 2007 Edition. http://www.wcoomd.org/ie/En/Topics_Issues/HarmonizedSystem/DocumentDB/TABLE_OF_CONTENTS_2007.html

[ISO2] International Standard Organization (ISO). "Codes for the representation of names of countries and their subdivisions." ISO 3166-1 ALPHA-2: 1997 (E/F), International Organization for Standardization. Geneva, 1997 (2006).

[ISO3] International Standard Organization (ISO): ISO 3166 ALPHA-3, 2006.

[ISSCAAP99] FAO. International Standard Statistical Classification of Aquatic Animals and Plants (ISSCAAP). Version in use until 1999 available at: ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/AnnexS1listISSCAAPold.pdf

[ISSCAAP00] FAO. International Standard Statistical Classification of Aquatic Animals and Plants (ISSCAAP). Version in use from 2000 available at: ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/AnnexS2listISSCAAP2000.pdf

[ISSCFVgrt] International Standard Statistical Classification of fishery Vessels (ISSCFV) by GRT Categories. in use until 1995. ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/annexL1ISSCFVgrt.pdf

[ISSCFV] International Standard Statistical Classification of Fishery Vessels (ISSCFV) by Vessel Types, in use until 1995. 1984. ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/annexLII.pdf

[ISSCFG] International Standard Statistical Classification of Fishing Gear (ISSCFG) ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/AnnexM1fishinggear.pdf

[ISSCFC] FAO. International Standard Statistical Classification of Fishery Commodities: Divisions and Group. ftp://ftp.fao.org/FI/DOCUMENT/cwp/handbook/annex/ANNEX_RII.pdf

[KIM09] Integrating country-based heterogeneous data at the United Nations: FAO's geopolitical ontology and services. S. Kim, M. Iglesias Sucasas, C. Caracciolo, V. Viollier, J. Keizer. Semantic Technology Conference 2009. Available at: http://www.semanticuniverse.com/articles-integrating-country-based-heterogeneous-data-united-nations-fao%E2%80%99s-geopolitical-ontology-and

[MB05] A. Miles and D. Brickley. SKOS Core Vocabulary Specification. Technical report, World Wide Web Consor tium (W3C), November 2005. http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102

[M49] United Nations. UN Code (M49). http://unstats.un.org/unsd/methods/m49/m49alpha.htm.

[OCEANATL] UN Ocean Atlas. http://www.oceansatlas.org/servlet/CDSServlet?status=ND1maWdpczE0Nzg3JjY9ZW4mMzM9KiYzNz1rb3M~

[ONEF] One fish topic tree. http://www.onefish.org/global/index.jsp

[ONTO101] N. F. Noy and D. L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[PGG] D. Picca, A. Gangemi, A. Gliozzo. LMM: an OWL Metamodel to Represent Heterogeneous Lexical Knowledge. Proc. of the International Conference on Language Resources and Evaluation (LREC) Marrakech, Morocco, 2008.

[PER05] C. Perez and S. Conrad. Relational.OWL - A Data and Schema Representation Format Based on OWL. In Proc. of APCCM 2005.

[RT] FAO. Table Selector for Reference Tables. http://www.fao.org/figis/servlet/RefServlet

[SITC3] United Nations Statistics Division. Standard International Trade Classification, Revision 3. http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=28&Lg=1

[SW] SemanticWorks. http://www.altova.com/

[TBC] TopBraid Composer. http://www.topbraidcomposer.com/

[XMLSpy] Altova. XMLSpy. http://www.altova.com/products/xmlspy/xml_editor.html