
**NeOn: Lifecycle Support for Networked Ontologies**

**Integrated Project (IST-2005-027595)**

**Priority: IST-2004-2.4.7 – "Semantic-based knowledge and content systems"**

# D7.2.2 Revised and Enhanced Fisheries Ontologies

**Deliverable Co-ordinator:** **Caterina Caracciolo**

**Author:** **Caterina Caracciolo**

**Contributor:** **Aldo Gangemi**

**Deliverable Co-ordinating Institution:**

**Food and Agriculture Organization of the United Nations (FAO)**

| | | | |
|---|---|---|---|
| Document Identifier: | NEON/2007/D7.2.2/v1.2 | Date due: | August 31, 2007 |
| Class Deliverable: | NEON EU-IST-2005-027595 | Submission date: | August 31, 2007 |
| Project start date: | March 1, 2006 | Version: | 1.2 |
| 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**<br>Knowledge Media Institute – KMi<br>Berrill Building, Walton Hall<br>Milton Keynes, MK7 6AA<br>United Kingdom<br>Contact person: Martin Dzbor, Enrico Motta<br>E-mail address: {m.dzbor, e.motta} @open.ac.uk | **Universität Karlsruhe – TH (UKARL)**<br>Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB<br>Englerstrasse 28<br>D-76128 Karlsruhe, Germany<br>Contact person: Peter Haase<br>E-mail address: pha@aifb.uni-karlsruhe.de |
| **Universidad Politécnica de Madrid (UPM)**<br>Campus de Montegancedo<br>28660 Boadilla del Monte<br>Spain<br>Contact person: Asunción Gómez Pérez<br>E-mail address: asun@fi.upm.es | **Software AG (SAG)**<br>Uhlandstrasse 12<br>64297 Darmstadt<br>Germany<br>Contact person: Walter Waterfeld<br>E-mail address: walter.waterfeld@softwareag.com |
| **Intelligent Software Components S.A. (ISOCO)**<br>Calle de Pedro de Valdivia 10<br>28006 Madrid<br>Spain<br>Contact person: Jesús Contreras<br>E-mail address: jcontreras@isoco.com | **Institut 'Jožef Stefan' (JSI)**<br>Jamova 39<br>SI-1000 Ljubljana<br>Slovenia<br>Contact person: Marko Grobelnik<br>E-mail address: marko.grobelnik@ijs.si |
| **Institut National de Recherche en Informatique et en Automatique (INRIA)**<br>ZIRST – 655 avenue de l'Europe<br>Montbonnot Saint Martin<br>38334 Saint-Ismier<br>France<br>Contact person: Jérôme Euzenat<br>E-mail address: jerome.euzenat@inrialpes.fr | **University of Sheffield (USFD)**<br>Dept. of Computer Science<br>Regent Court<br>211 Portobello street<br>S14DP Sheffield<br>United Kingdom<br>Contact person: Hamish Cunningham<br>E-mail address: hamish@dcs.shef.ac.uk |
| **Universität Koblenz-Landau (UKO-LD)**<br>Universitätsstrasse 1<br>56070 Koblenz<br>Germany<br>Contact person: Steffen Staab<br>E-mail address: staab@uni-koblenz.de | **Consiglio Nazionale delle Ricerche (CNR)**<br>Institute of cognitive sciences and technologies<br>Via S. Martino della Battaglia,<br>44 - 00185 Roma-Lazio, Italy<br>Contact person: Aldo Gangemi<br>E-mail address: aldo.gangemi@istc.cnr.it |
| **Ontoprise GmbH. (ONTO)**<br>Amalienbadstr. 36<br>(Raumfabrik 29)<br>76227 Karlsruhe<br>Germany<br>Contact person: Jürgen Angele<br>E-mail address: angele@ontoprise.de | **Food and Agriculture Organization of the United Nations (FAO)**<br>Viale delle Terme di Caracalla 1<br>00100 Rome<br>Italy<br>Contact person: Marta Iglesias<br>E-mail address: marta.iglesias@fao.org |
| **Atos Origin S.A. (ATOS)**<br>Calle de Albarracín, 25<br>28037 Madrid<br>Spain<br>Contact person: Tomás Pariente Lobo<br>E-mail address: tomas.parientelobo@atosorigin.com | **Laboratorios KIN, S.A. (KIN)**<br>C/Ciudad de Granada, 123<br>08018 Barcelona<br>Spain<br>Contact person: Antonio López<br>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 writing parts of this document:

UPM

CNR

USFD

## Change Log

| Version | Date | Amended by | Changes |
|---|---|---|---|
| 0.1 | 10-07-2007 | Caterina Caracciolo | First Draft |
| 0.2 | 31-07-2007 | Caterina Caracciolo | Completed description of domains and ontology models. Added discussions, lessons learned. |
| 1.0 | 17-08-2007 | Caterina Caracciolo | Revision of the entire document. |
| 1.1 | 19-09-2007 | Caterina Caracciolo | Implemented comments from QA |
| 1.2 | 20-09-07 | Andrew Bagdanov,<br>Caterina Caracciolo | Linguistic revision.<br>Added Annex IV provided by Aldo Gangemi, harmonized with the rest of the document |

## 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/aos/fi. This document is organized as follows. In Chapter 1 we place our work in the context of the WP7 case study. In Chapter 2 we describe previous attempts to create ontologies for the fisheries domain. In Chapter 3 we recap the user requirements presented in our previous deliverable D7.1.1, with special attention to the functionalities needed for modelling, population and maintenance. In Chapter 4 we describe the domains and the data on which the ontologies described here are based. In Chapter 5 we describe the fisheries database where the data used to populate the fisheries ontologies is stored; we also introduce the tool used for population of the ontologies. In Chapter 6 we describe the models of all ontologies produced. In Chapter 7 we discuss some features of the ontologies. In Chapter 8 we summarize the lessons learned in the course of this work. Finally, in Chapter 9 we draw our conclusions. This document also includes four Annexes: the list of naming conventions adopted (Annex I), an essential glossary of fisheries terms (Annex II) a list of acronyms (Annex III) and a report on the conversion of the XML schema for fisheries factsheets into an ontology.

# Table of Contents

## List of tables



## Lit 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 will use FAO and non-FAO datasets on fisheries and it will access them by means of a network of ontologies.

Deliverable D7.1.1 [D7.1.1] presented the user requirements for the WP7 use case, including both the FSDAS and the lifecycle for the ontologies to be used in it. The user groups involved in the lifecycle were analyzed and special attention was paid to the requirements for ontology editors (i.e., domain experts, translators, and information management specialists) in charge of the daily maintenance of the ontologies. Visualization and editing facilities for these users were especially recommended. The next deliverable produced in WP7, Deliverable D7.2.1 [D7.2.1], was dedicated to an inventory of electronic resources available for the fisheries domain. In that deliverable, a number of resources were described, and 28 of them were selected for inclusion in the FSDAS system based on their suitability and importance. From these resources, we selected the reference data used for the collection and dissemination of fisheries statistics. Classification codes stored in the reference data are also used in the creation of XML fact sheets for fisheries. Reference data is stored in a relational database.

Relational databases are very efficient for storing and retrieving parent-child hierarchies. However, fisheries reference data is stored as a set of separate hierarchies and despite the existence of documents and knowledge evidencing relationships among elements of the various hierarchies (e.g. a biological species living in a particular sea, caught and exported in the form of different commodities). Data sitting within the various hierarchies lacks integration and relationships are neither implemented nor managed. The problem is then how to include relations across the various domains in order to allow the fisheries community around the world to better analyze the current state of fisheries and evaluate trends measuring the impact of various factors. The goal is to fully exploit the knowledge stored in the reference data to include relations across domains while keeping all functionalities currently supported.

Our approach to this issue consists in adding a semantic layer on top of the database by modelling as ontologies the domains covered by the reference data while keeping the actual data instances in the database. In this way any number of additional relationships can be handled at the ontology level, while at the same time the efficiency and all the functionalities provided by the RDBMS are kept. This solution would also allow legacy systems connected to the database to continue to work without additional effort. In order to successfully implement this solution, we need to be able to map complex database schemas to ontologies and to efficiently access the data in the database both in batch mode and at run-time. As an intermediate step, we experiment with the “static” generation of ontologies populated with data from DBs. The following two steps (run-time access to the database and the addition of relations at the ontology level) are left for future work.

Each ontology described in this deliverable is thus a “stand-alone” ontology that covers one domain (e.g., fishing areas, fisheries commodities etc.) of reference data. These ontologies are populated with data *extracted* from the database according to an ontology model created on the basis of the domain at hand and the structure of the database. This is done in order to:

1. verify that it is possible to correctly access all data from the database (i.e., no relevant piece of data in the database should remain inaccessible),
2. verify that it is possible to perform the extraction according to a sound ontological model, and

3. produce ontologies that can be used by WP7 and other NeOn partners in the next phases of the project.

However, in real life applications, it would be preferable to leave the data in its physical location (database) and access it through a semantic layer (ontologies) at run time. The reasons for preferring this setting are the following. First, since data sets can be large, it is convenient to exploit the efficiency of an RDBMS and to spare expensive and error prone processes of data conversion. Second, by keeping data in their current location, all applications that access the data will continue to work.

In deliverable D7.4.1 [D7.4.1] the three major steps in the fisheries ontologies lifecycle are described and depicted in a figure (Chapter 2) that is reproduced below (Figure 1). This figure summarizes the steps involved in the lifecycle, together with the actions that take place in each step and the user groups involved. The ontologies presented in the current deliverable did not go through the entire lifecycle, as they went from step 1 (iteration of conceptualization and population) directly to step 3 (publishing in the production environment). The intermediate step 2 (validation and update) was skipped for the twofold reason that the focus of our work was on the extraction of data instances from a relational database according to an ontological format, and because of the lack of ready-to-use tools for enabling ontology editors to validate and if necessary, to update the ontologies.

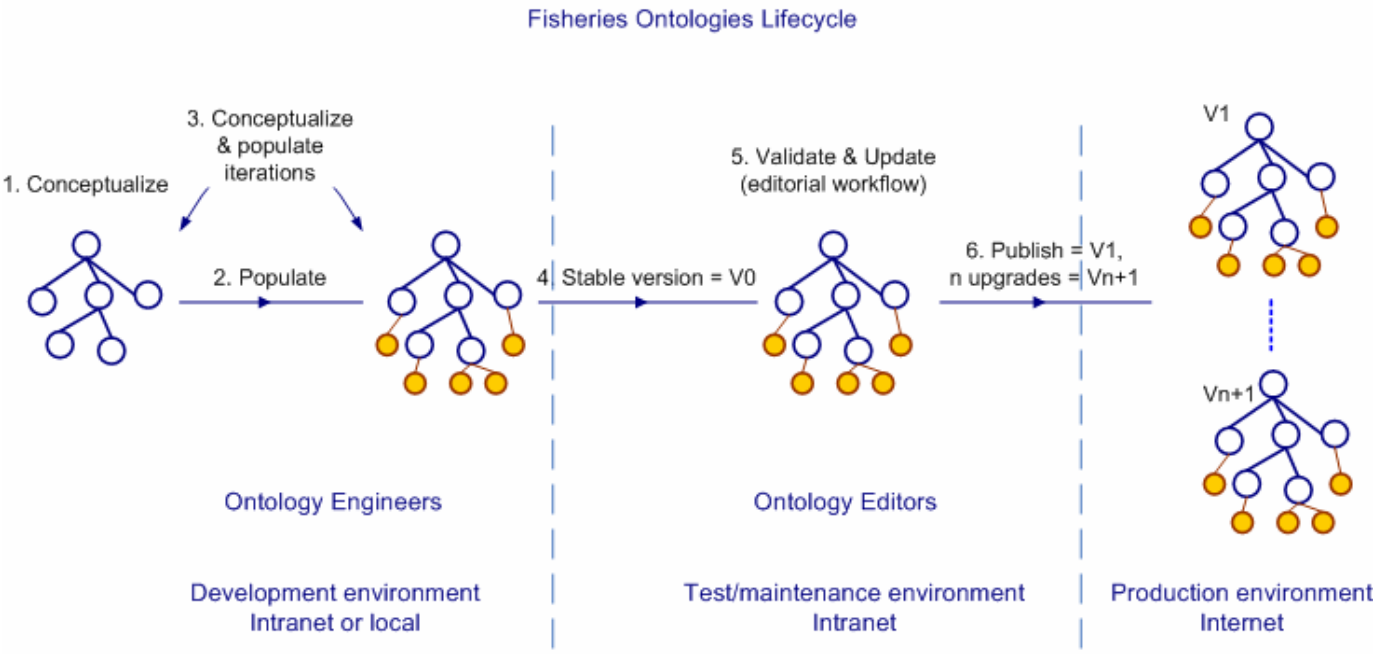


**Figure 1. Major steps in the fisheries ontologies lifecycle (figure from D7.4.1, Chapter 2).**

Summarizing, the ontologies presented in this deliverable constitute the first set of fisheries ontologies produced within WP7. They will be improved in a second phase of the project[1] in several ways: in terms of the way data is accessed in the database, in the amount and type of information they contain, and with links between ontologies in order to form a network.

The rest of this deliverable is organized in the following way. In Chapter 2 we survey the Fishery Ontology project (FOS), a predecessor of the NeOn project for fisheries. In Chapter 3 we provide a brief recap of requirements for tools for managing the lifecycle of ontologies in WP7. In Chapter 4 we describe the domains covered by the reference data. In Chapter 5 we describe in more detail

[1] The result of that phase will be consolidated in deliverable D7.2.3: “Enhanced networked fisheries ontologies”, due at month 30.

the structure of the database of reference data, and the tool we used to populate the ontologies from the database. The models of the ontologies created are described in detail in Chapter 6 (ontologies are published on the FAO website and so made publicly available). In Chapter 7 we highlight and discuss some relevant aspects of the ontologies produced. In Chapter 8 we summarize the lessons learned while carrying out our task, and in Chapter 9 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. Finally, in Annex IV we report on an exploratory study on the conversion of the XML schema for fisheries factsheets into an ontology.

# 2 Previous work: the FOS project

The Fishery Ontology Project (FOS), in operation from 2002 to 2003, was managed jointly by ISTC-CNR and FAO [GAN04WW]. It was designed for "the creation, integration and utilization of ontologies for information integration and semantic interoperability in fisheries information systems." A detailed analysis of the FOS project can be found in deliverable D7.1.1, Annex I. Here we recap the salient aspects of that project and highlight the lessons learned from it.

At that time, integration and interoperability were interpreted as having one centralized, consistent library of ontologies that played the role of a *hub* helping the interoperability between different document servers or other information systems. The approach adopted in FOS consisted of the following steps:

1. reengineering informal or semi-structured terminological and metadata resources (KOSs: knowledge organization systems) into formal ones;
2. organizing and aligning the reengineered KOSs within an appropriate layered and modular *ontology library*.

The FOS project used the following resources (details provided refer to the time the project was carried out):

1. **OneFish topic trees [ONEF]** is a hierarchy of topics with average depth of three, organized into five disjoint categories called 'worldviews' (subjects, ecosystems, geography, species, administration), plus one worldview (stakeholder) maintained by the users of the community. Topics listed under more than one parent are marked with @.
2. **AGROVOC [AGROVOC]** is the thesaurus used in FAO to index documents related to all areas of interest of the Organization. The fragment of AGROVOC related to fisheries was used in FOS, consisting of approximately 2,000 fisheries related descriptors (out of 16,000 descriptors). The fragment was manually extracted;
3. **ASFA thesaurus [ASFA]** is the thesaurus used to index the Aquatic Science and Fisheries Abstracts (ASFA) collection of documents, which covers the world's literature on the science, technology, management, and conservation of marine, brackish water, and freshwater resources and environments, including their socio-economic and legal aspects.
4. **FIGIS reference tables [RT]** is the dataset used by FAO to reference statistical data on fisheries. At the time of the project, it included approximately 200 top-level concepts, with maximal depth of 4. It also contains 30,000 'objects' (mixed concepts and individuals), relations (specialized for each top category, but scarcely instantiated) and multilingual support.

The approach followed in FOS embodied a three-layered structure: a *foundational layer*, a *core layer* and a *domain layer*. The idea of this three-layered structure is that each top-class or property in domain ontology is a subclass of a class resp. property in the core ontology, and each top-class or property in the core ontology is a subclass of a class resp. property in the foundational ontology. For example, Yellow Tuna (domain ontology) rdfs:subClassOf Biological entity (core ontology), which rdfs:subClassOf Physical entity (foundational ontology).

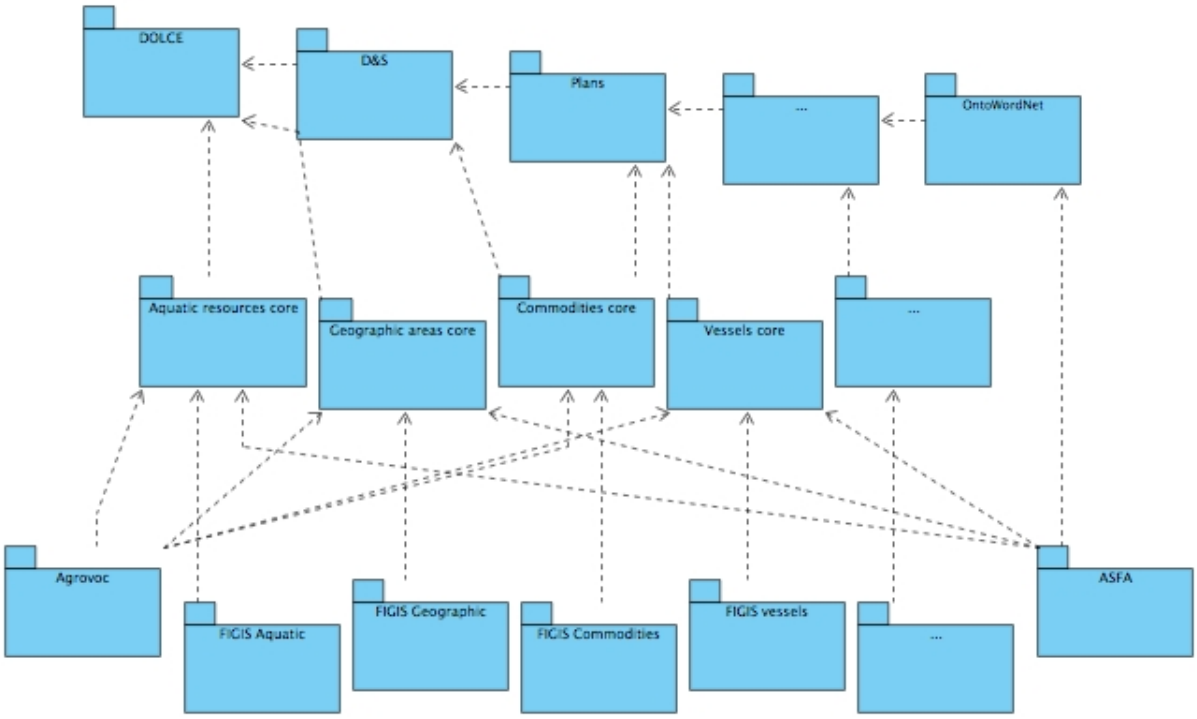


**Figure 2. A diagram representing the three-layered structure of FOS.**

Figure I depicts the dependencies between some of the ontologies from the three layers of the FOS library. The topmost layer contains the reused *foundational* ontologies. The middle layer shows the *core* ontologies of fisheries, the bottom layer shows some *domain* fisheries ontologies that contain dependencies to the core ones.

After the completion of FOS, the resources on which the domain ontologies were built continued to evolve, while the fisheries ontologies remained static. This happened because FOS ontologies did not enter the editorial cycle used by the people who maintained the original resources and were barely integrated into the network of resources of the organization. Moreover, since there was no mechanism to automatically reflect updates from the sources to the ontologies (and back), it was natural for editors to keep working on the sources and avoid duplicating addition/change of material. Summarizing, we identified the following limitations in FOS:

1. lack of an automatic way to maintain the mapping between FOS ontologies and the underlying resources when updated. In fact, the mappings were stored in conversion tables without an implemented mechanism to maintain the mappings over the dynamics of both ontologies and sources;

2. lack of integrated tools to create the domain ontologies by starting from the original resources. A formal workflow was defined, but not implemented in a set of integrated, smoothly-working tools;

3. lack of automatic methods to extract “modules” from the original resources and create the ontologies.

The lessons learned from the FOS project where then integrated in the user requirements for the ontology lifecycle gathered in deliverable D7.1.1. We summarize the relevant requirements for the present deliverable in the next section.

# 3 Requirements on modelling, population and maintenance

In this section we recap the requirements introduced in Deliverable D7.1.1 [D7.1.1] concerning the tools needed for the lifecycle of fisheries ontologies. We distinguish two groups of requirements: those concerning modelling and population of ontologies and those concerning their maintenance. The former group of activities is typically performed by ontology engineers (cf. [D7.1.1] Section 4.3), the latter by ontology editors (cf. [D7.1.1] Section 4.4).

## 3.1 Modelling and population

The entire process of ontology modelling should be supported. This includes functionalities for:

1. visualizing ontology elements (i.e., classes, datatype and object properties, URIs and metadata) of one or more ontologies at a time. The visualized ontologies may or may not be connected in a network;
2. visualizing mappings between ontologies and between ontologies and underlying resources such as databases;
3. editing all ontology elements of one or more ontologies at a time;
4. automatically creating and updating documentation concerning the main features of the ontology (e.g., names of all ontology elements including mappings, metadata and summarizing statistics);
5. linking to relational databases in order to allow the creation and population of ontologies based on data residing in them.

We stress the fact that an appropriate visualization is important both for population by manual editing and for population by connection to databases.

## 3.2 Maintenance

Tools for ontology maintenance are crucial to keeping the ontologies up-to-date and so ensuring that they can be used in real applications. It is therefore important that:

1. in case of manual maintenance (editing), all ontology elements can be manually edited and visualized, and that the appropriate metadata (e.g., author and date of the change) is stored together with its history;
2. in case of automatic population from a database, the database should be automatically checked for updates and a list of changes presented to the editor;
3. ontology versioning be supported;
4. mapping between ontologies be supported and all functionalities of editing, visualization and versioning available to it;
5. ontologies can be part of a workflow. In particular, more than one editor should be allowed to work (i.e. edit and visualize) on the same ontology, if possible with constraints (permissions) imposed by user profiles and associated with entire ontologies and modules in it.

# 4 Reference data

In the inventory presented in deliverable D7.2.1, 28 systems including both FAO and non FAO resources, were carefully detailed. Based on that analysis, we selected the FAO resources at the core of many information systems in fisheries: the reference data used to collect, store and access statistical data and to produce XML fact sheets on fisheries.

The FAO Fisheries and Aquaculture Information and Statistics Service (FIES) collates statistics concerning several aspects of fisheries. A time series is a sequence of observations which are ordered in time and/or space. FIES collects observations about captures, aquaculture production, catches, fleets, trade of commodities, and consumption [FISTAT]. Each 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). In the case of statistics concerning trade, also the "trade flow" (import/export) is included. Table 1 provides an example of statistics relative to import and export of "tunas, skipjack and Atlantic bonito, prepared or preserved" in the year 2000.

| Land Area | Trade flow | Commodity | 2000 |
|---|---|---|---|
| Algeria | Export | Tunas, skipjack and Atlantic bonito, prepared or preserved | 1 |
| | Import | Tunas, skipjack and Atlantic bonito, prepared or preserved | 841 |
| **Total Algeria** | | | 842 |
| **Grand total** | | | 842 |

**Table 1. Import and export of fisheries commodities in Algeria in the year 2000.**

The data used to indicate the dimensions are called reference data and are organized into Reference Tables (RT) [RT]. Reference tables store the *codes* assigned to reference data according to one or more coding system maintained by international organizations. They also store 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). 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).[2] The entire system that manages the RT is called Rerefence Tables Management System (RTMS), whose core is an Oracle database, called FIGIS.

Reference data is also used in the fisheries fact sheets [FS] where a large amount of information about fisheries, aquaculture and related subjects, including fishing techniques, fishing areas, fisheries and aquaculture country profiles, is made available to the public in the form of semi-structured text. All fisheries fact sheets in FAO are in XML format, structured according to a comprehensive XML schema [FSschema] that includes all elements used in all types of fact sheets. Fact sheets are organized 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.

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

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 ontologicial model, but was rather organised following hierarchical document formatting conventions. A dictionary of the elements used in the schema is available online [FSdic].

In the rest of this chapter we describe in detail each hierarchy of reference data used to generate the ontologies described in Chapter 6.

## 4.1 Land areas

Most fisheries statistics are on production and catch and are reported by individual countries. Data can then be aggregated above the national level into groups defined according to different criteria, such as 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.[3]

Codes used for land areas are the ISO-3166 ALPHA-2 [ISO2] and ALPHA-3 [ISO3] codes maintained by the International Standard Organization (ISO), and the M49 [M49] code maintained by the UN Statistical Division.

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.

Since territories and groups change over time, the RT also includes their range of validity in order to continue to be able to query the statistical database according to territories no longer existing. For example, territories can 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 economical groups similarly change every time a country joins or leaves a group.

## 4.2 Fishing areas

Marine and inland waters are divided into regions, or "FAO division areas," for the purpose of data collection and statistical reporting (for the entire list of FAO divisions, see [FAOdiv]). The FAO division areas consist of major areas, divided into sub-areas, each divided into divisions, and these finally into sub-divisions. This 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 FAO code used for these areas is a taxonomic code.[4] For example, the major area “SouthEast Pacific” has code 87, and one of its subdivisions has code: 87.2.1.1. Figure 3 represents the FAO major Fishing area 51, Western Indian Ocean, and its subareas, numbered from 1 to 8.

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

[4] 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.

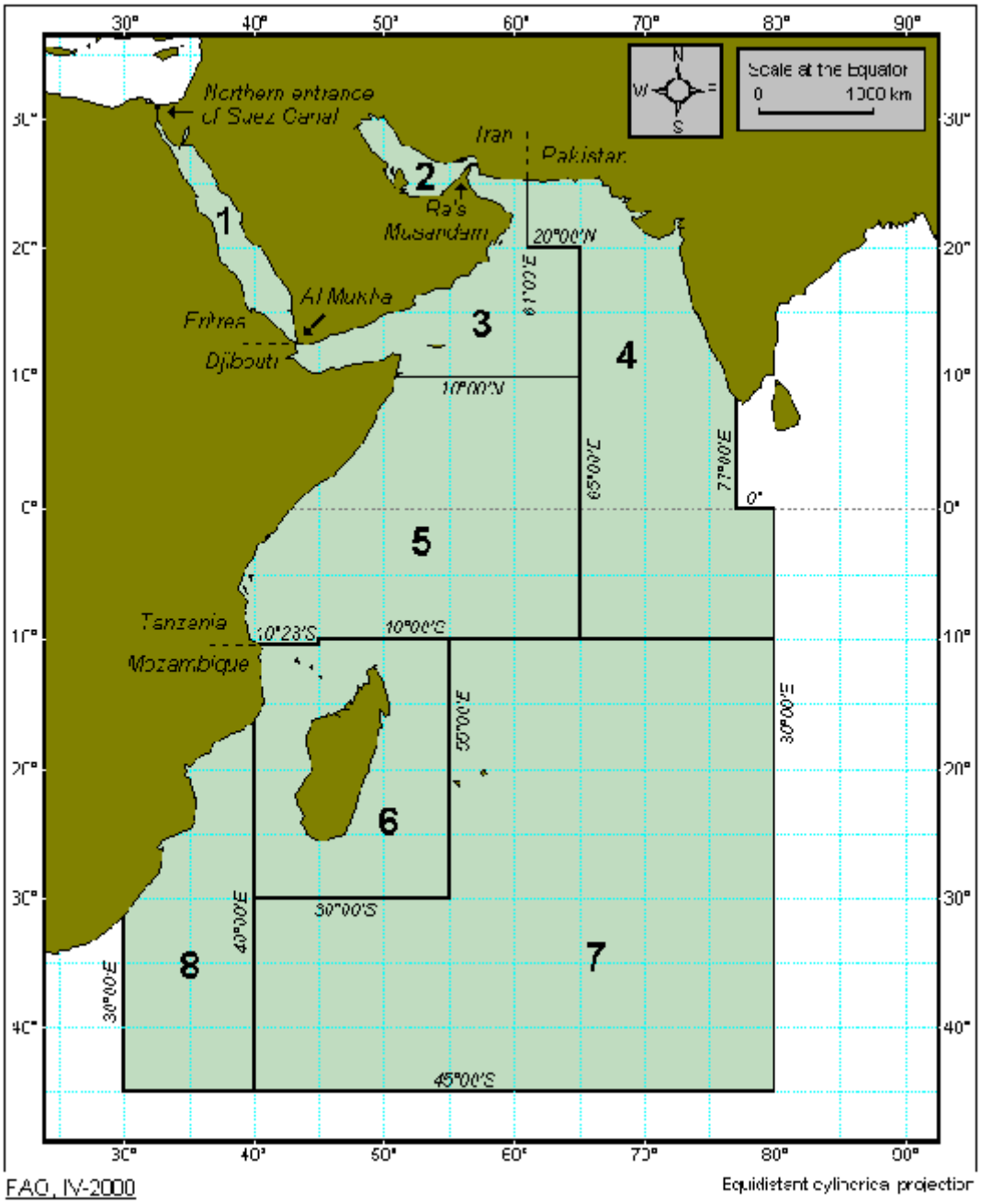


**Figure 3. An example of FAO major fishing area: Western Indian Ocean (FAO code 51).**

## 4.3 Biological entities

Reference data about biological species is used for a number of statistics, such as catch, production and trade. These statistics are collated at either the species or higher taxonomic levels. Each level is referred to as a *species items*. Species items are organized and maintained in the Aquatic Science and Fisheries Information System (ASFIS). For each species item a taxonomic code and the ISO ALPHA-3 [ISO3] are provided. An English name is available for most of the

records, and about one third of them have also a French and Spanish name.[5] Currently, the ASFIS list includes nearly 11.000 species items selected according to their interest or relation to fisheries and aquaculture.

A taxonomic code for biological entities is a 10-digit code 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 2 analyzes an example of taxonomic code (of a biological species: 1750400301) to show how the 10 digits of the taxonomic code are organized (the family "above" that species has taxonomic code: 17504XXXXX).

| | *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 2. Structure of the 10-digit taxonomic code used for biological entities.**

Since a biological entity (i.e., main group, order, family, genus, species) is only included in the reference table if/when there is data associated with it, the taxonomic classification included in the database may not be complete (i.e., there can be species for which only the main group they belong to is specified, and no order nor family are given).

The 3-alpha code identifier is a unique code made of three letters that is widely used for the exchange of data with national correspondents and among fisheries agencies.

## 4.4 Fisheries commodities

Fisheries commodities cover products derived from any aquatic animal (fish, crustaceans, molluscs) and residues caught for commercial, industrial or subsistence uses, by all types of classes of fishing units operating in inland, fresh and brackish waters, in inshore, offshore or high seas fishing areas.

Several coding and classification systems are available for fisheries commodities. FAO's International Standard Statistical Classification of Fishery Commodities (ISSCFC) [ISSCFC] is used for detailed information on countries or zones. The ISSCFC is an expansion of the United Nations Standard International Trade Classification (SITC) [SITC3], developed by the United Nations' Statistical Office on the basis of earlier international work on the subject. The ISSCFC is linked with the Harmonized Commodity Description and Coding System (abbreviated to HS) [HS07] of the World Customs Organization (WCO).[6] The ISSCAAP classification is also used (see section above).

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. The Harmonized System (HS) is

[5] 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.

[6] 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 of about 139 countries and territories.

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.

The SITC coding system reflects various aspects of commodities 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.

## 4.5 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 engaged in international voyages.

Although the London Convention has been adopted for vessels of 24 metres in length and over, for many vessels only data conforming to the Oslo Convention are available. The situation varies from country to country.[7]

Based on the international convention in use, FAO fleet data on the vessel tonnage are 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.

[7] 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).

## 4.6 Gear types

The type of gear installed on a vessel determines the type of fish that it can catch, therefore it is often used in statistical collection 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 harmonised 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 factsheets.

# 5 Creation and population of ontologies from the FIGIS database

In this section we briefly describe the structure of the database in which the reference data is stored in order to point out the salient features that need to be taken into consideration when populating ontologies with data coming from such database.

## 5.1 The FIGIS database

The main idea underlying the database of reference tables is that anything is an *item* that can have information attached. For example, a territory is an item (endowed with one or more names, codes according to various international coding systems and so on) that can be represented as a row in a table. Similarly, a continent is also thought of as an individual (also endowed with names and codes) represented as a row in the same table. According to the same scheme, all elements of a biological taxonomy (e.g., species, orders) are considered items.

A flag (called *meta code*) is used to distinguish what *type* of object each item is, e.g., a country, a species, a water subdivision and so on. Meta codes are organized into a strict hierarchy, with a common root (called *figis object* after the name of the database) under which each domain is shaped as a strict sub-hierarchy. For example, all domains described in the previous section correspond to a sub-hierarchy starting at the first level of the main hierarchy. Table 3 presents a fragment of the FIGIS hierarchy of meta codes.

1 figis object
- 10 000 Land area
  - 11 000 Geographical region
    - **11 002** Continents
  - 12 000 Groups and union of countries
    - **12 001** Economic unions
  - 13 000 Country, political or statistical entity
    - **13 001** Country
- 20 000 Water area
  - 21 000 Environmental area
    - **21 001** Inland/Marine
  - 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
- 30 000 Biological entity
  - 31 000 Taxonomic entity
    - **30 001** Group
    - **30 002** Order
    - **30 003** Family
    - **30 005** Species
- 40 000 Fishery commodities
  - **40 001** Commodity (FAO ISSCFC classification)
  - **40 002** Commodity (Harmonized classification)
- 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 3. A fragment of the hierarchy of meta codes (those used are in bold).**

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: [8]

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 etc.), and
2. one *group table*, in which the actual hierarchy is stored. The group table is a (four-column) table that renders any hierarchy as a group-member structure. It contains the ID from the item table of both group and member (foreign keys), plus the meta code of the group.

Note that there is only one meta table in the database, so 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, and the item and group tables corresponding at the domain at hand. For example, in order to get and interpret all reference data concerning biological species (cf. Section 4.3) 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 4).

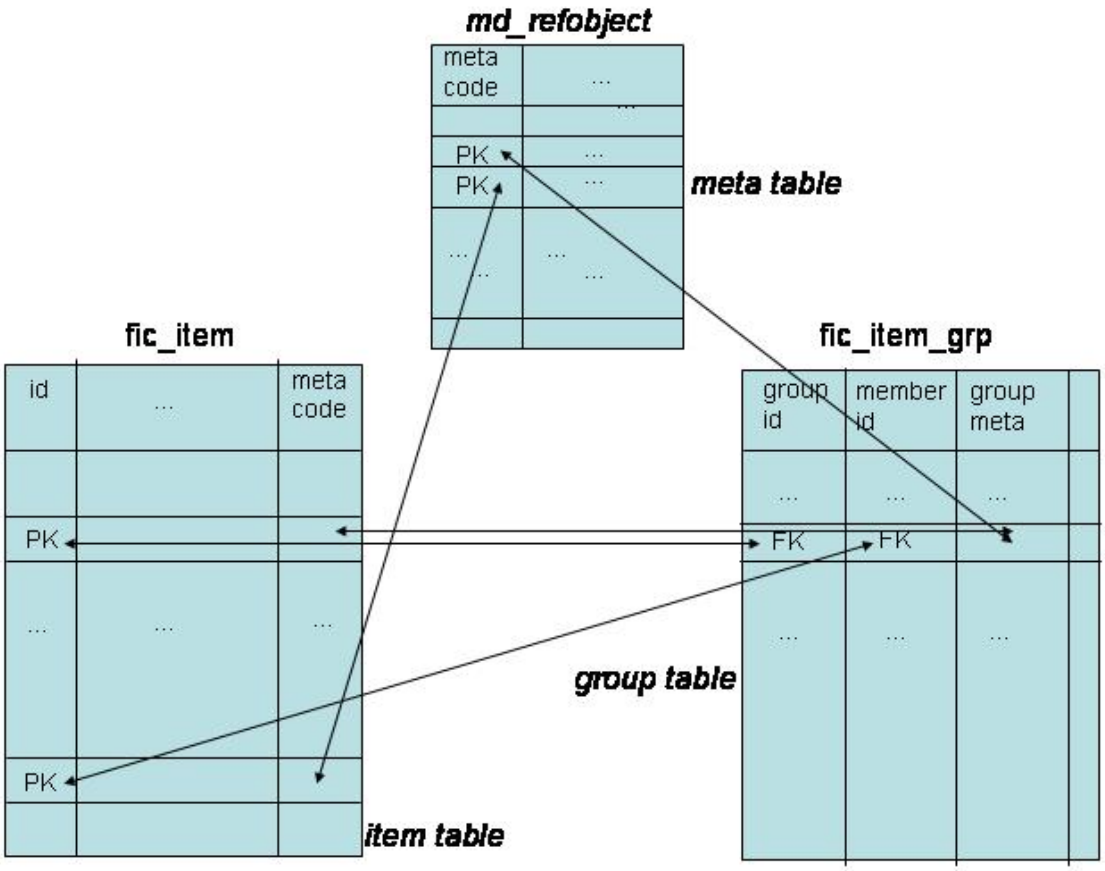


**Figure 4. The FIGIS database: tables for the domain of biological entities.**

In some cases, a special type 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

[8] 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.

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.

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 5).

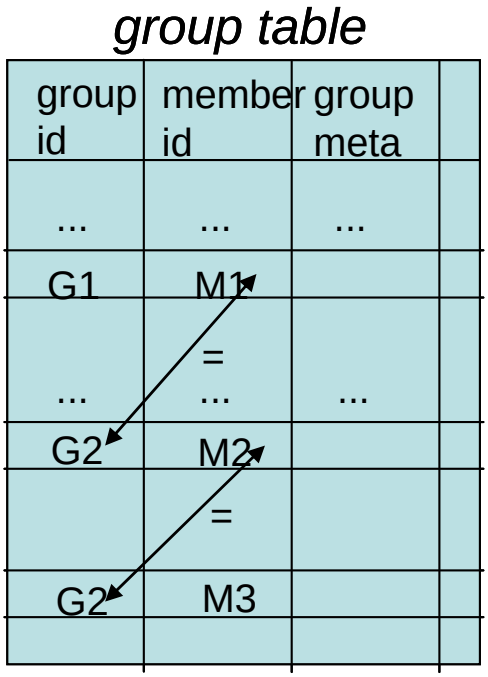


**Figure 5. Typical structure of a group table, where an element appears both as group and member (e.g., M1=G2).**

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.

## 5.2 Population of ontologies from database

Various techniques exist for populating ontologies from existing databases, including [BIZ03, BAR03, PER05]. For our task we selected ODEMapster [BAR06,BAR07], a tool currently under development at the Universidad Politecnica de Madrid.[9]

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 complex 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 of 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 process that generates all possible ontology individuals from the data repository. The operations of ODEMapster are not limited by the

[9] The NeOn toolkit currently allows one to query a database on the basis of an ontology, but does not allow the export of entire data sets according to an ontology model.

expressivity of the DBMS. The set of primitives can be extended with delegable or non delegable primitive conditions and operations. The processor will delegate the execution of certain actions to the RDBMS and execute 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.

## 5.3 Iteration of conceptualization and population

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. Below we describe the two main phases of the process we followed, namely domain conceptualization and the actual population of the ontologies, and highlight the iterative nature of this process (cf. step 1, Figure 1, Chapter 1).

### 5.3.1 Conceptualization

In order to obtain an adequate knowledge of the domain covered by the reference data, we studied all the available material (bibliographic references are provided in Chapter 4), including the relevant fact sheets from the Handbook of Fishery Statistical Standards by the Coordinating Working Party on Fishery Statistics (CWP), and the actual classification systems used. 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,

The first models for the ontologies based on reference data were created only by looking at the domain, as explained above. We used Protege 3.1.1 to create and edit the ontology models (on the basis of common metholodogies and best practices for ontology creation, such as [ONTO101]). In many cases we created two alternative models for each domain: the main differences between these models concerned the modelling of codes and names in various languages (i.e., datatype properties vs object properties) and that of hierarchies (e.g., subclasses of biological entities, and part-of hierarchies of land areas). We wrote the corresponding documentation in which we analyzed the pros and cons of each choice. It was decided that the preference for one model over another should be determined not only on the basis of a sound modelling, but also of the efficiency in actual use of the ontology, and of the efficiency in getting the data to populate the ontology.

### 5.3.2 Population

In order to analyze the FIGIS database we integrated the study of the available documentation (cf. Section 5.1) with a number of interviews with the information experts working with the database. From these interviews we obtained a deeper understanding of the FIGIS 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.

From the analysis of the database we found that different flavors of hierarchies are present and encoded in a similar way (cf. Section 7.3). We also found that often only a subset of the available classification system is used, as in the case of biological entities, which taxonomy does not include the Genus (because no timeseries are available for that), or as in the case of fisheries commodities. Fisheries commodities are also an example of hierarchical classifications that are

stored in a non-hierarchical way in the database. These findings made us adjust the ontology models in order to accommodate the specificities of the available data.

### 5.3.3 Iteration of modelling and population

We used ODEMapster (cf. Section 5.2) to automatically populate the ontology. The ontologies presented in this deliverable are the result of a number of iterations of the cycle conceptualization-population. In the course of these iterations, a number of modelling decisions were made (for a detailed discussion see Chapter 7) for the twofold reason of improving the quality of the models and making them suitable for use together with ODEMapster to extract data from the database. This was the case, for example, for the decision between datatype and object properties, being the former type being easier to extract than the latter.