A Formal Ontological Framework for Semantic Interoperability in the Fishery Domain

Aldo Gangemi[1], Frehiwot Fisseha[2], Ian Pettman[3], Domenico M. Pisanelli[1], Marc Taconet[4], Johannes Keizer[2]

[1] Institute of Psychology, CNR (National Research Council), Rome, Italy
{gangemi,pisanelli}@ip.rm.cnr.it
http://saussure.irmkant.rm.cnr.it
[2] FAO-GILW, Rome, Italy
{Frehiwot.Fisseha,Johannes.Keizer}@fao.org
http://www.fao.org
[3] One Fish, SIFAR, Grange-over-Sands, Cumbria, UK
ip@ceh.ac.uk
http://www.onefish.org[3]
[4] FIDI, FAO, Rome, Italy
Marc.Taconet@fao.org
http://www.fao.org

Abstract. This paper outlines a project (involving FAO, SIFAR, and CNR) aimed at building an ontology in the fishery domain. The ontology will support semantic interoperability among existing fishery information systems and will enhance information extraction and text marking, envisaging a fishery semantic web. The ontology is being built through the conceptual integration and merging of existing fishery terminologies, thesauri, reference tables, and topic trees. Integration and merging are shown to benefit from the methods and tools of formal ontology.

1 INTRODUCTION

1.1 The general problem

Specialized distributed systems are the reality of today's information systems architecture. Developing specialized information systems/resources in response to specific user needs and/or area of specialization has its own advantage in fulfilling the information needs of target users. However, such systems usually use different knowledge organization tools such as vocabularies, taxonomies and classification systems to manage and organize information. Although the practice of using knowledge organization tools to support document tagging (thesaurus-based indexing) and information retrieval (thesaurus-based search) improves the functions of a particular information system, it is leading to the problem of integrating information from different sources due to lack of semantic interoperability that exists among knowledge organization tools used in different information systems.

The different fishery information systems and portals that provide access to fishery information resources are one example of such scenario. This paper demonstrates the proposed solution to solve the problem of information integration in fishery information systems. The proposal shows how a fishery ontology that integrates the different thesauri and taxonomies in the fishery domain could help in integrating information from different sources be it for a simple one-access portal or a sophisticated web services application.

1.2 The local scenario

Fishery Ontology Service (FOS) is a key feature of the Enhanced Online Multilingual Fishery Thesaurus, a project aimed at information integration in the fishery domain. It undertakes the problem of accessing and/or integrating fishery information that is already partly accessible from dedicated portals and other web services.

The organisations involved in the project are: FAO Fisheries Department (FIGIS), ASFA Secretariat, FAO WAICENT (GIL), the oneFish service of SIFAR, and the Ontology and Conceptual Modelling Group at ISTC-CNR. The systems to be integrated are: the "reference tables" underlying the FIGIS portal [1], the ASFA online thesaurus [2], the fishery part of the AGROVOC online thesaurus [3], and the oneFish community directory [4].

The official task of the project is "to achieve better indexing and retrieval of information, and increased interaction and knowledge sharing within the fishery community". The focus is therefore on tasks (indexing, retrieval, and sharing of mainly documentary resources) that involve recognising an *internal structure* in the content of texts (documents, web sites, etc.). Within the semantic web community and the intelligent information integration research area (cf. [5] and [6]), it is becoming widely accepted that content capturing, integration, and management require the development of detailed, formal *ontologies*.

In this paper we sketch an outline of the FOS development and some hint of the functionalities that it carries out.

2 ONTOLOGY INTEGRATION AND MERGING

2.1 Heterogeneous systems give heterogenous interpretations

An example of how formal ontologies can be relevant for fishery information services is shown by the information that someone could get if interested in *aquaculture*.

In fact, beyond simple keyword-based searching, searches based on tagged content or sophisticated natural-language techniques require some conceptual structuring of the linguistic content of texts. The four systems concerned by this project provide this structure in very different ways and with different conceptual

'textures'. For example, the AGROVOC and ASFA thesauri put *aquaculture* in the context of different thesaurus hierarchies; an excerpt of the AGROVOC result is (*uf* means *used for*, *NT* means *narrower than*; *rt* means *related term, Fr* and *Es* are the corresponding French and Spanish terms):

AQUACULTURE
uf aquiculture
uf mariculture
uf sea ranching
NT1 fish culture
NT2 fish feeding
NT1 frog culture

rt agripisciculture
rt aquaculture equipment

Fr aquaculture
Es acuicultura

The AGROVOC thesaurus seems to frame aquaculture from the viewpoint of *techniques* and *species*. On the other hand, the ASFA aquaculture hierarchy is substantially different:

AQUACULTURE
uf Aquaculture industry
uf Aquatic agriculture
uf Aquiculture
NT Brackishwater aquaculture
NT Freshwater aquaculture
NT Marine aquaculture
rt Aquaculture development
rt Aquaculture economics
rt Aquaculture engineering
rt Aquaculture facilities

Actually this hierarchy seems to stress the *environment* and *disciplines* related to aquaculture.

A different resource is constituted by the so-called *reference tables* in FIGIS system; the only reference table mentioning *aquaculture* puts it into another context (taxonomical species):

Biological entity
Taxonomic entity
Major group
Order
Family
Genus
Species
Capture species (filter)
Aquaculture species (filter)
Production species (filter)

Tuna atlas spec

The last resource examined is oneFish directory, which returns the following context (related to *economics* and *planning*):

SUBJECT
Aquaculture
Aquaculture development
Aquaculture economics @
Aquaculture planning

With such different interpretations of *aquaculture*, we can reasonably expect different search and indexing results. Nevertheless, our approach to information integration and ontology building is not that of creating a homogeneous system in the sense of a reduced freedom of interpretation, but in the sense of navigating alternative interpretations, querying alternative systems, and conceiving alternative contexts of use.

To do this, we require a comprehensive set of ontologies that are designed in a way that admits the existence of many possible pathways among concepts under a common conceptual framework. This framework should reuse domain-independent components, be flexible enough, and be focused on the main reasoning schemas for the domain at hand.

Domain-independent, *upper* ontologies should characterise all the general notions needed to talk about economics, biological species, fish production techniques; for example: *parts*, *agents*, *attribute*, *aggregates*, *activities*, *plans*, *devices*, *species*, *regions of space or time*, etc. While the so-called *core* ontologies should characterise the main conceptual habits (schemas) that fishery people actually use, namely that certain plans govern certain activities involving certain devices applied to the capturing or production of a certain fish species in certain areas of water regions, etc.

Upper and core ontologies [7,8] provide the framework to integrate in a meaningful and *intersubjective* way different views on the same domain, such as those represented by the queries that can be done to an information system.

2.2 Methods applied to develop the integrated fishery ontology

Once made clear that different fishery information systems provide different views on the domain, we directly enter the paradigm of *ontology integration*, namely the integration of schemas that are arbitrary logical theories, and hence can have multiple models (as opposed to database schemas that have only one model) [9]. As a matter of fact, thesauri, topic trees and reference tables used in the systems to be integrated could be considered as *informal* schemas conceived to query semi-formal or informal databases such as texts and tagged documents.

In order to benefit from the ontology integration framework, we must transform informal schemas into *formal* ones. In other words, thesauri and other terminology management resources must be transformed into (formal) ontologies.

To perform this task, we apply the techniques of three methodologies: OntoClean [8], ONIONS [10], and OnTopic [11].

The first one contains principles for building and using upper ontologies for core and domain ontology analysis, revision, and development. In its current form, OntoClean also features an axiomatised domain-independent top-level of formal criteria, concepts and relations (Figure 3) [18].

ONIONS is a set of methods for enhancing the informal data of terminological resources to the status of formal ontological data types. Some methods are aimed at reusing the structure of hierarchies (e.g., BT/NT relations, subtopic relation, etc.), the additional relations that can be found (e.g., RT relations), and at analysing the compositional structure of terms in order to capture new relations and definitional elements. Other methods concern the management of semantic mismatches between alternative or overlapping ontologies, and the exploitation of systematic polysemy to discover relevant domain conceptual structures.

OnTopic is about creating dependencies between topic hierarchies and ontologies. It contains methods for deriving the elements of an ontology that describe a given topic, and methods to build 'active'topics that are defined according to the dependency of any individual, concept, or relation in an ontology.

In Figure 1, a class diagram is shown of the informal and formal data types taken into account by the forementioned methodologies.

In section 3.1 the types of (meta)data extracted from the resources are described. In the subsequent sections the (meta)data types obtained from the transformation of resources into a merged ontology are also described.

We briefly describe:

- the resources that are integrated
- how the Integrated Fishery Ontology (IFO) is being built
- a mediation architecture to interface the fishery ontology service with the source information systems.

3 OUTLINE OF THE FOS PROJECT

3.1 Resources

The following resources have been singled out from the fishery information systems considered in the project:

the **oneFish** topic trees (about 1,800 topics), made up of *hierarchical topics* with brief summaries, identity codes and attached knowledge objects (documents, web sites, various metadata). The hierarchy (average depth: 3) is ordered by (at least) two different relations: *subtopic*, and *intersection between topics*, the last being notated with @, similarly to relations found in known subject directories like DMOZ. There is one 'backbone' tree consisting of five disjoint categories, called *worldviews* (*subjects, ecosystem, geography, species, administration*) and one worldview (*stakeholder*), maintained by the users of the community, containing own topics and topics that are also contained in the first four other categories (Figure 5). Alternative trees contain new 'conjunct' topics deriving from theintersection of topics belonging to different categories.

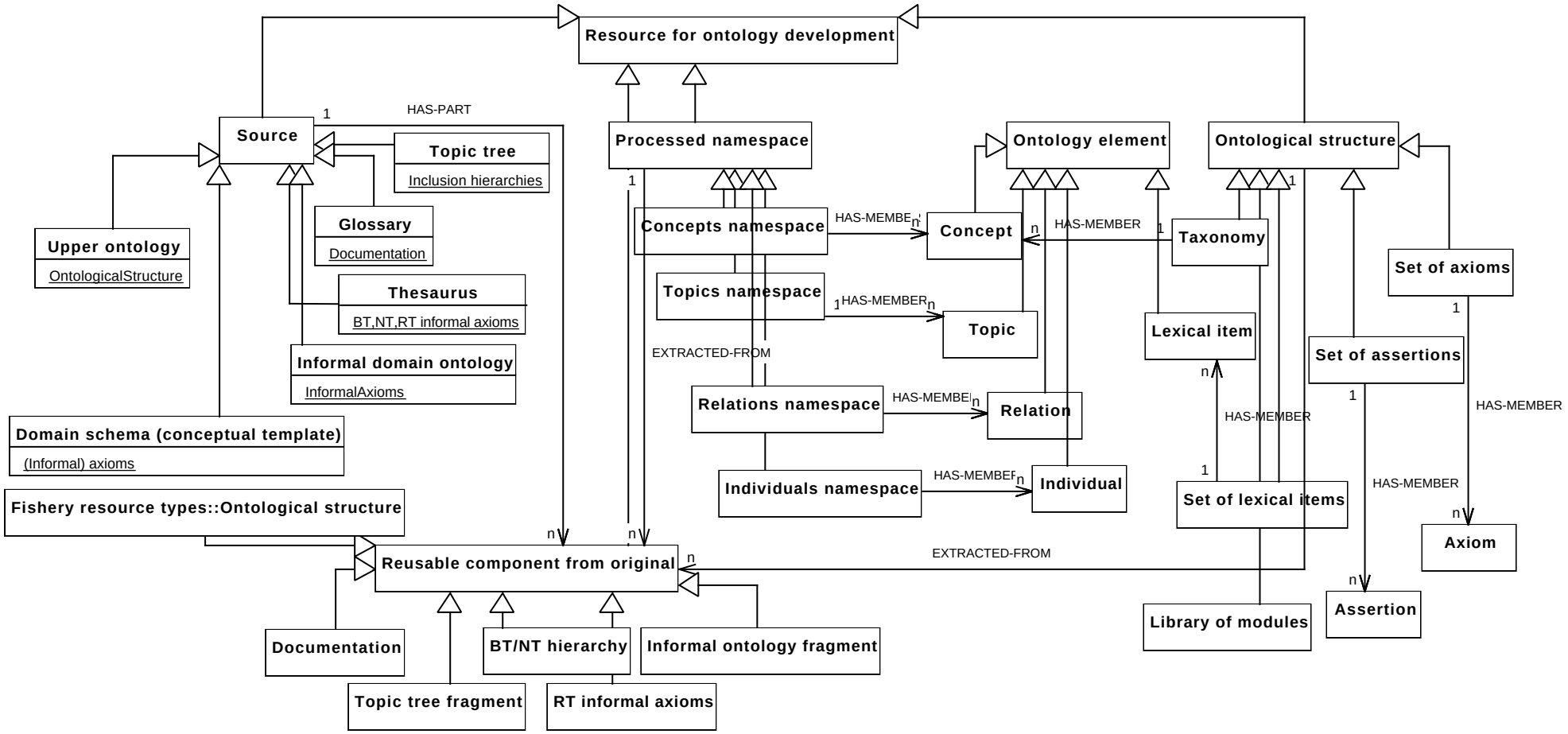

Fig. 1. A class diagram of the source data types taken into account

AGROVOC thesaurus (about 500 fishery-related descriptors), with thesaurus relations (*narrower term*, *related term*, *used for*) among descriptors, lexical relations among terms, terminological multilingual equivalents, and glosses (*scope notes*) for some of them.

ASFA thesaurus, similar to AGROVOC, but with about 10,000 descriptors.

FIGIS reference tables, with 100 to 200 top-level concepts, with a max depth of 4, and about 30,000 'objects' (mixed concepts and individuals), relations (specialised for each top category, but scarcely instantiated) and multilingual support. There are modules (*water areas*, *continental areas*, *biological entities*, *vessels*, *commodities*, *stocks*, etc.), also organised by 'views'.

In Figure 2 a diagram is sketched of the methodology used to extract and refine the informal data from the fishery information systems. The methodology is also described in the next sections.

3.2 Translation and refining of the components for IFO building

The (meta)data from the resources that have been singled out have been processed, in order to integrate them within a homogeneous environment, and with a clear assessment of their nature. In the following we list a set of guidelines that have been followed to translate and refine data components:

- A detailed evaluation of each source (find the schema -explicit or not- underlying the implementation of source data, then describe each data type both qualitatively and quantitatively) is performed.
- A language to represent the KB is chosen that hosts the integration activity. A description logic like DLR [9] is an ideal choice for its compatibility with the ontology integration framework.
- An ontology server is installed that supports DLR or compatible languages.
- Some data types from the sources (Figure 1) seem appropriate to be included in a preliminary prototype. The following steps are performed on them:
    - Discuss, refine and formalise FIGIS fishery conceptual schemas [12] to build a preliminary core ontology. Also the upper-level concepts from the source thesauri should be matched against the FIGIS conceptual schemas. This results in a *resource for core ontology development*.
    - Translate FIGIS reference tables: taxonomy, individuals, and local relations (to be transformed into formal axioms). This results in a *draft resource for domain ontology development*.
    - Reuse oneFish topic trees to design a preliminary architecture for IFO library. This architecture should match the preliminary core ontology. This results in a *resource for ontology library design*.

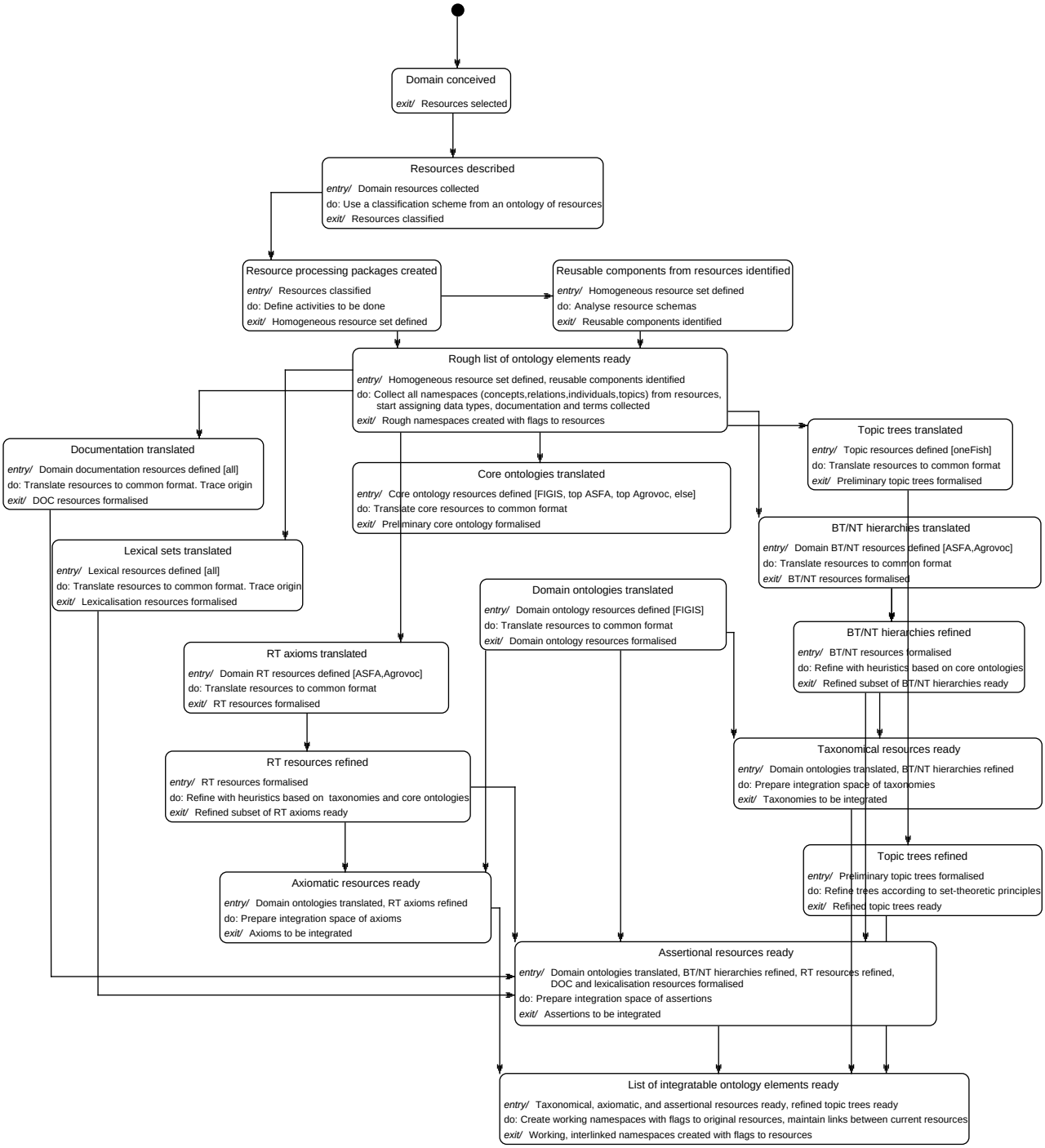

Fig. 2. A diagram of the methodology used to extract and refine the informal data

- Extract IS_A taxonomies from AGROVOC and ASFA BT/NT (*Broader Term/Narrower Term*) hierarchies. Heuristics from upper and core ontologies can be applied to clean up BT/NT hierarchies, for example, the following rule can be applied: *if a body part descriptor is NT of an organism descriptor, then this is probably* not *an IS_A use of NT* (probably it is a *part-of* relation). This results in *resources for core and domain taxonomies building*.
- Expand RT (*Related Term*) relations from AGROVOC and ASFA. Also non-IS_A BT/NT hierarchies could be refined (expanded) here. Heuristics can be applied here as well, for example, *if there exists a systematic relation between to concepts in the core ontology, and there exists a RT relations between two subconcepts of those concepts, then this is an indication for that relation to be the refinement of the RT one*. This results in *resources for core and domain axioms building*.
- Reuse UF (*Used For*) relations and (multi-)linguistic equivalents from all resources. Track must be kept of the context from which a linguistic item has been extracted. This results in *resources for ontology lexicalisation*.

3.3 Parallel tasks

In the following sections we outline the main steps to build the basic taxonomy, documentation, and architecture for the integrated fishery ontology.

3.3.1 Developing a fishery core ontology (FCO)

In this step, we pick up uppermost concepts and conceptual (categorisation) schemas from sources and integrate them with a certified top-level containing domain-independent concepts, relations and meta-properties. The resources needed for such a task are:

Upper ontology resources: the OntoClean upper level [8,18] (Figure 3) is a preferential choice for its compatibility with the methodology. For alternatives, see [13]. Moreover, various formal ontologies and standards for relations, and general lexical repositories like WordNet [14].

Core ontology resources: conceptual templates, (selected in the preliminary phases), relational database schemas, theoretical views on domain topics, domain standards, etc. An informal fishery core ontology (the FIGIS *composite concepts*) is shown in Figure 4.

In the context of core ontology development, some taxonomical branches (*core concepts*) have relevant conceptual integration issues that are being studied by ontological engineers and domain experts in close collaboration:

- *biological taxonomies*: difficult having a stable framework of reference (in principle, mapping from local taxonomies to a biological one is feasible, but in practice it could be not cost effective)
- *geographic regions*: use GIS as a stable framework of reference? geographic names?
- *institutions*: maybe automatic clustering of individuals through classification

- *fishing devices* (including vessels)
- *fishing and fish farming techniques* (plans and activity types)
- *farming systems* (sets of components)
- *fishery regulations* (norms)
- *fishery managament systems* (plans)
- *production centers*

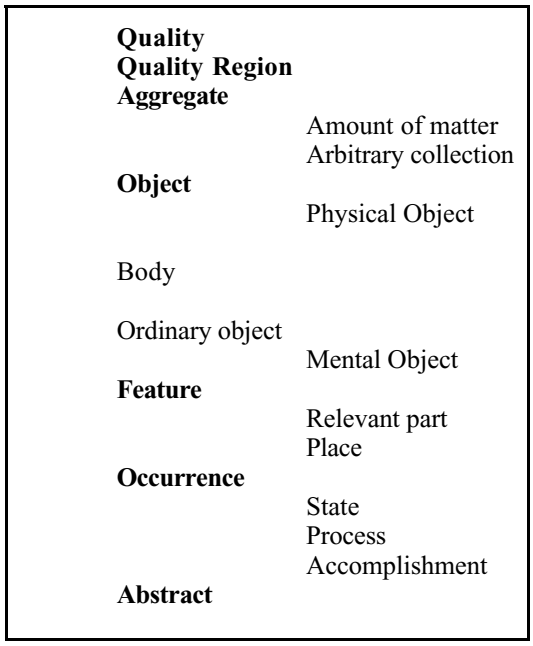

Fig. 3. The OntoClean top concepts

Development is performed as incremental loading and classification of upper and core level ontologies in the Ontology Server.

Another indirect resource that can be exploited to build the core ontology is the analysis of *systematic polysemies* (they have been already used in the mining of large medical thesauri, cf.[10]). A systematic polysemy is discovered when a relation exists between two senses of a term, and this relation is *relevant* for the domain that is being analysed. Consequently, if we find many polysemies with senses that have been conceptualised within the same concept pairs, this is an indication for a possible core ontology relation.

3.3.2 Building domain IS-A taxonomies

This phase deals with the integration of the resources for domain ontology development with the fishery core ontology (developed in the previous phase).

Resulting taxonomies could be either 'tolerated' or 'cleaned up'. Tolerance amounts to have widespread and unexplained polysemy for terms, but it is not time consuming. Cleaning is the most time consuming task, since a frequent scenario is the following: concept C from source S1 (C^S1) is in principle similar to a D^S2

(usually because they share one or more terms), but they actually occupy two taxonomical places that make them disjoint according to the upper or core ontology.

The ONIONS methodology [10] in this case suggests to axiomatise their glosses (cf. ⁄3.2.3, 3.3.3) and to check if their taxonomical position is correct. If it is not, then they are probably polysemous senses of the same term, and some alternative methods can be applied to relate those senses, to merge them, or to accept the conceptual split of the senses.

Some cleaning will be needed in any case to remove at least the major taxonomical clashes. This results into a *domain taxonomy*. Additional effort should be dedicated to distinguish:

Concepts vs individuals (heuristics applicable: country names, institutions, etc.).

Backbone concepts vs viewpoint concepts (roles, reified properties, contingent notions), cf. [7,8].

This eventually results into a refined domain taxonomy.

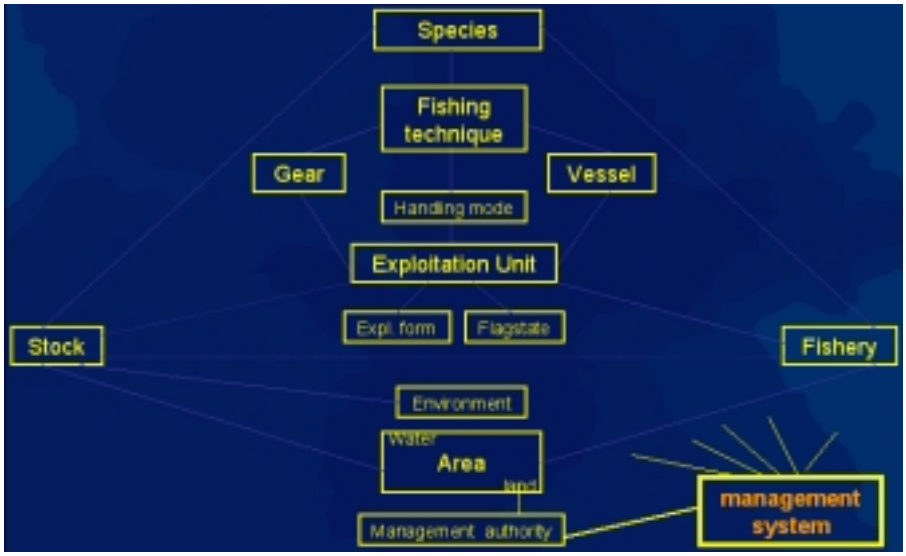

Fig. 4. The FIGIS *composite concepts*, used as a resource for core ontology development.

3.3.3 Collecting existing documentation and producing glosses

Available resources for ontology documentation are collected and associated as a kind of annotation (*gloss*) to domain concepts. Concepts lacking a gloss require a new one.

For core concepts and relations, besides existing glosses, an extensive description of their scope in the FCO is provided.

3.3.4 Designing a preliminary topic architecture

A preliminary topology for most general topics (to be used for ontology modularisation as well) is figured out. Here the following resources are reused: ontologies for topics (Welty s topic topology [15], topic maps standard [16], OnTopic principles [11]), semantic portals design [17], *oneFish* topic trees.

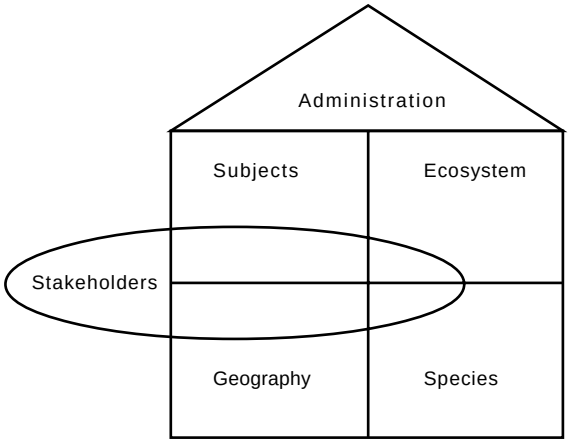

Fig. 5. Topic spaces (*"worldviews"*) in oneFish.

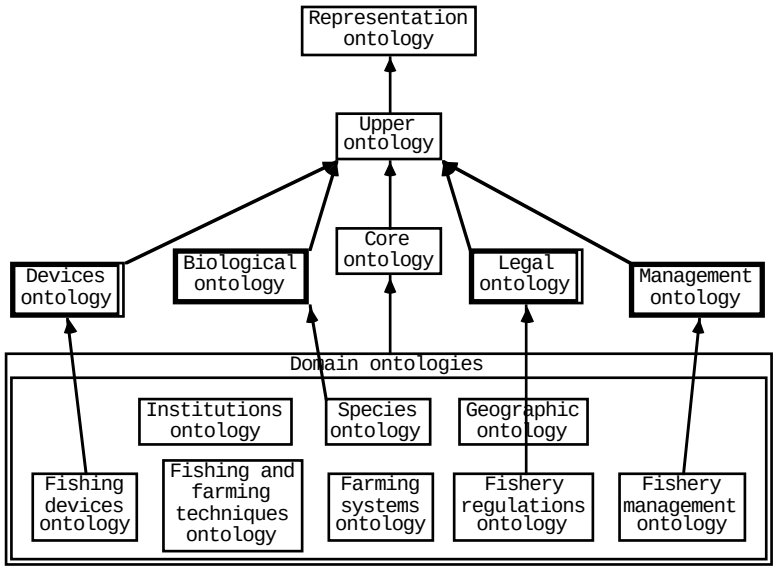

Fig. 6. An example architecture for the fishery ontology library. Double frames mean external ontologies.

The topic topology will be used both for maintaining the ontology library and for managing text indexing and retrieval. Figure 5 shows how the current topic spaces of oneFish are structured. Figure 6 shows an ontology-based architecture for the Integrated Fishery Ontology.

3.4 Building domain axioms

Once taxonomies are cleaned to a certain extent, documented, and divided into appropriate namespaces, activities aimed at raising the conceptual detail of the ontology can be started. The most important is the characterisation of domain concepts with axioms. In order to realise this, domain resources containing informal relationships, and (at least some) glosses from documentation are upgraded to the status of logical axioms.

Informal relationships can be found in thesauri (e.g. *related term*) as well as reference tables and topic trees. They are mined in order to understand:

1) if the axioms are applicable to all the subconcepts of the concept to which the axiom pertain, and
2) what quantification is applicable to those axioms: existential (necessary) or universal (contingent)?

This results into formal Domain Axioms. This axiom set is enhanced by axiomatising glosses. Here the ONIONS methodology [10] is applied to derive formal domain axioms from natural language descriptions. The typical technique consists in extracting terms, parsing them according to a dependency grammar, and applying core and upper ontologies to assign concepts and relations to the resulting dependency trees.

This activity is time-consuming, and semi-automatic techniques are still a research issue [13]. Scalability and approximate results are considered here.

The axioms obtained from informal relationships and glosses are revised according to the fishery core ontology developed so far.

3.5 Modularising ontology library according to topics

Following OnTopic methodology [11], dependency chains of core concepts are automatically generated and the existing preliminary topic topology is checked in order to produce a first version of the ontology library architecture. Dependency chains are also applied to derive indexing tags and boolean search spaces.

A dependency chain is the transitive closure of the logical depend-ons of a concept. The transitive closure is applied to the defining elements of a concept. Here a set of *relevance parameters* are applied in order to

3.6 Providing multi-lingual lexicalisation to elements in the ontology library

An integrated fishery ontology benefits from the existence of terms already related to concepts in the original resources, since they semi-automatically provide the so-called *lexicalisation* of concepts. On the other hand, having an integrated ontology also provides a powerful tool to check polysemous senses of terms, as well as to check consistency of UF thesaurus relations and consistency of multi-lingual equivalents.

3.7 A unified architecture

Figure 7 shows a simplified example architecture to support information brokering [6] or unified search after merging of fishery information systems by means of Fishery Ontology Service.

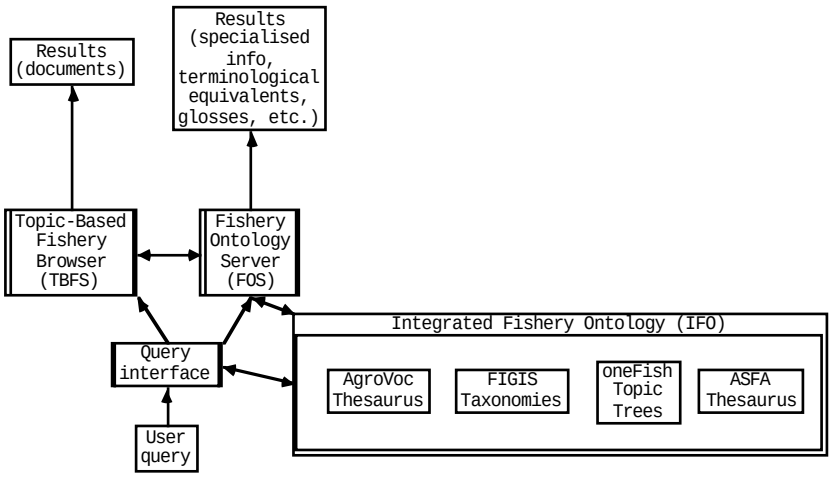

Fig. 7. A unified interface for interoperability after merging heterogeneous terminological resources in fishery.

The basic idea is that user queries, through a query interface, can be submitted to two kinds of servers: if the query aims at retrieving documents, a topic-based fishery agent rewrites the query in order to submit it to heterogeneous databases (*brokering*); if the query aims at finding specialised conceptual or terminological information, it is directed to the Fishery Ontology Server (FOS). In both cases, the query interface uses FOS. Query rewriting needs also mapping relations from the integrated fishery ontology to the source thesauri.

CONCLUSIONS

In this paper we have outlined some research solutions within the framework of ontology integration that are based on formal upper and core ontologies. Some details have been given on how informal schemata such as thesauri, reference tables, and topic trees can be reused and refined in order to be manipulated by ontology integration. Some hints have also been shown about the dependence of topic trees from ontologies, a promising research area for the semantic web.

In fact, the overall research issue underlying the FOS project is to provide a unified methodology of ontology integration and merging based on formal ontologies, ontology library design, topic trees building and maintainance, and efficient web search and indexing.

REFERENCES

[1] http://www.fao.org/fi
[2] http://www4.fao.org/asfa
[3] http://www.fao.org/agrovoc
[4] http://www.onefish.org
[5] http://www.ontoweb.org
[6] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/i3.html
[7] Gangemi A, Guarino N, Masolo C, Oltramari A.: Understanding Top-Level Ontological Distinctions, in: H. Stuckenschmidt (ed), *Proceedings of the IJCAI 2001 Workshop on Ontologies and Information Sharing* (2001)
[8] Gangemi A, Guarino N, Oltramari A.: Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level, in: C Welty, B Smith (eds.), *Proceedings of the 2001 Conference on Formal Ontology and Information Systems,* Amsterdam, IOS Press (2001)
[9] Calvanese D, De Giacomo G, Lenzerini M.: A Framework for Ontology Integration. Proceedings of 2001 Int. Semantic Web Working Symposium (SWWS 2001) (2001)
[10] Gangemi A, Pisanelli DM, Steve G.: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, 1999, vol.31, pp. 183-220 (1999)
[11] Gangemi A, Pisanelli DM, Steve G.: The OnTopic Methodology for Supporting Active Catalogues with Formal Ontologies. ISTC-CNR-OCMG Internal Report iii-01 (2001)
[12] Taconet M, Roux O: FIGIS, The Fisheries Global Information System.
[13] http://www.ontoweb.org/SIG
[14] Velardi P, Missikoff M, Fabriani P: Using Text Processing Techniques to Automatically Enrich a Domain Ontology, in: C Welty, B Smith (eds.), *Proceedings of the 2001 Conference on Formal Ontology and Information Systems,* Amsterdam, IOS Press (2001)
[15] Welty C, The Ontological Nature of Subject Taxonomies, N Guarino (ed.), *Proceedings of the First Conference on Formal Ontology and Information Systems,* Amsterdam, IOS Press (1998)
[16] Pepper S, The TAO of Topic Maps: http://www.gca.org/papers/xmleurope2000/papers/s11-01.html
[17] Stojanovic N, Maedche A, Staab S, Studer R, Sure Y: SEAL —A Framework for Developing SEmantic PortALs
[18] Oltramari A., Gangemi A, Guarino N, Masolo C,: Restructuring WordNet's Top-Level: The *OntoClean* approach, in K Simov (ed): *Proceedings of the The LREC2002 Workshop on Ontologies and Text*, Las Palmas (2002)