Building An Integrated Formal Ontology for Semantic Interoperability in the Fishery Domain

Aldo Gangemi[1], Frehiwot Fisseha[2], Ian Pettman[3], Johannes Keizer[2]

[1] Institute of Psychology, CNR (National Research Council), Rome, Italy
gangemi@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/agris
[3] One Fish, SIFAR, Grange-over-Sands, Cumbria, UK
ip@ceh.ac.uk
http://www.onefish.org

Abstract. This paper outlines a project (involving FAO, SIFAR, and CNR) aiming at building a formal 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 of existing fishery terminologies, thesauri, reference tables, and topic trees.

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 IP-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 Methodologies to support ontology integration

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 (with a penchant for kinds of *techniques* and *species*):

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

while the ASFA result is substantially different (it seems to stress the *environment* for aquaculture):

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
…

FIGIS reference tables may interpret *aquaculture* in still 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

and oneFish directory 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 be domain-independent, flexible enough, and focused on the main reasoning schemas for the domain at hand.

For example, the domain-independent (*'upper'*) ontologies should characterise all the general notions needed to talk about economics, biological species, fish production techniques; 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 kind in certain areas of water regions, etc.

Upper and core ontologies [7,8] provide the framework to integrate in a meaningful 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 said 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 [9]. As a matter of fact, the 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 data bases 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 deals with the use of upper ontologies and general principles for core and domain ontology building, the second describes several methods for enhancing the informal data of terminological resources to the status of formal ontology data types, the third shows how to create links between topic hierarchies and ontologies.

In Figure 1 a class diagram is shown of the informal and formal data types taken into account, while in Figure 2 a state diagram is sketched of the methodology used to extract and refine the informal data from the fishery information systems.

In the next section 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. Alternative trees contain new 'conjunct' topics deriving from the intersection of topics belonging to different categories.

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.

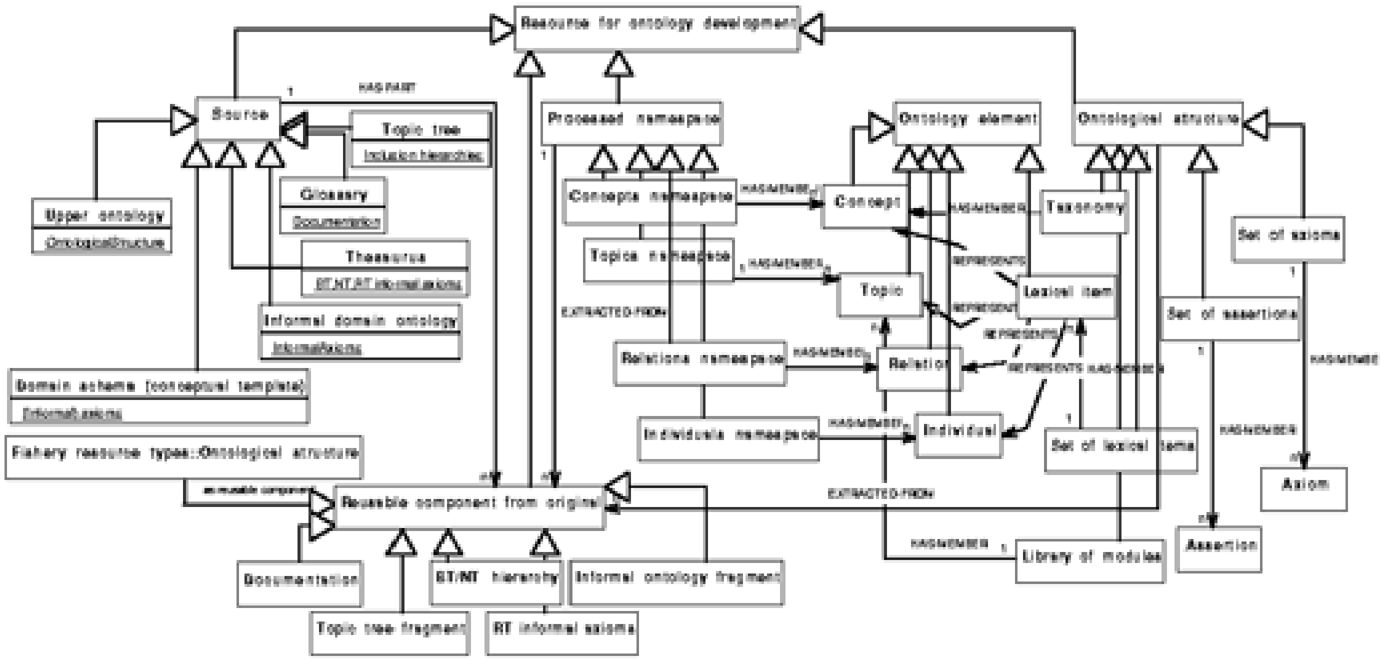

Fig. 1. A class diagram of the source data types taken into account.

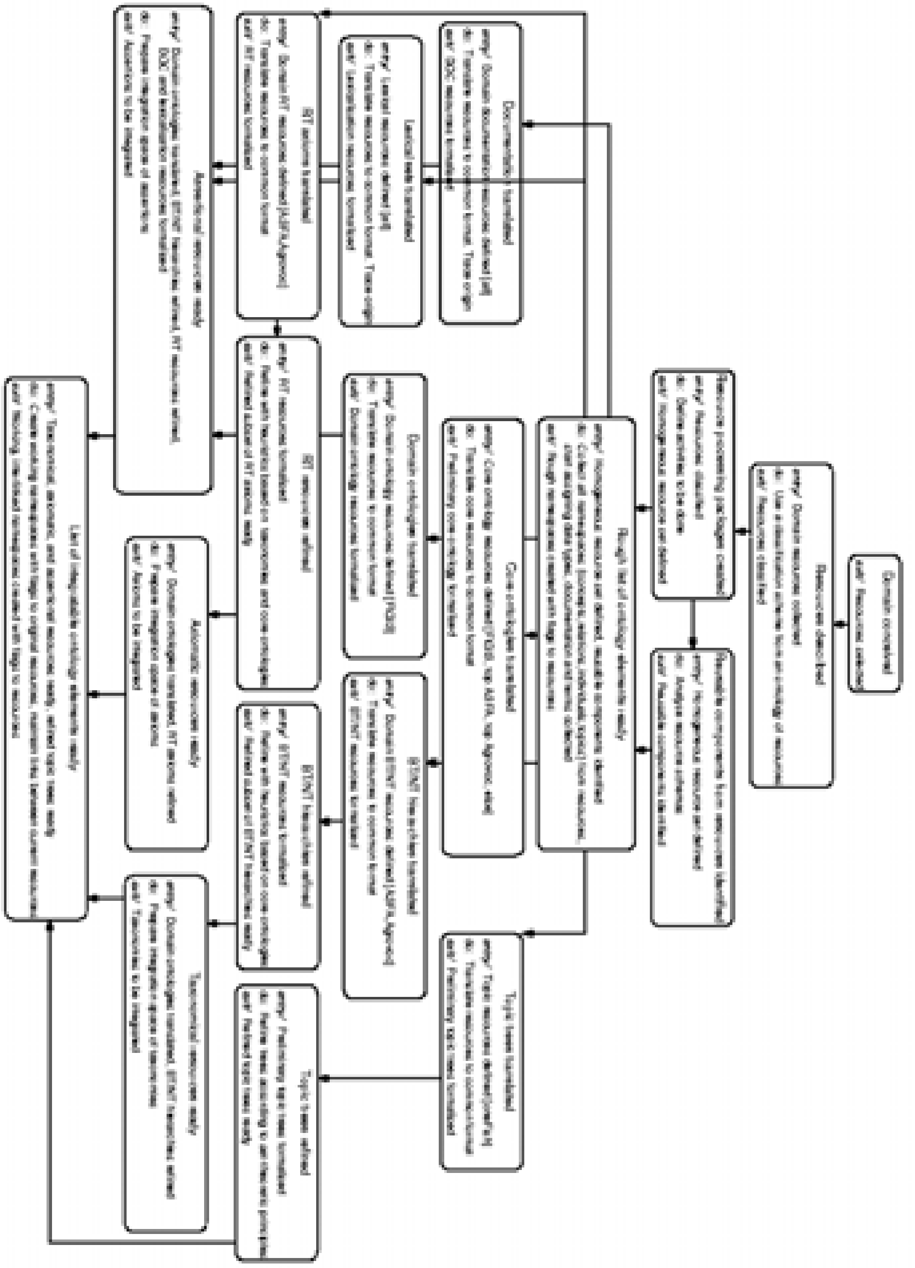
Domain conceived
exit/ Resources selected
Resources described
entry/ Domain resources collected
do: Use a classification scheme from an ontology of resources
exit/ Resources classified
Resource processing packages created
entry/ Resources classified
do: Define activities to be done
exit/ Homogeneous resource set defined
Reusable components from resources identified
entry/ Homogeneous resource set defined
do: Analyse resource schemas
exit/ Reusable components identified
Rough list of ontology elements ready
entry/ Homogeneous resource set defined, reusable components identified
do: Collect all namespaces (concepts, relations, individuals, topics) from resources, start assigning data types, documentation and terms collected
exit/ Rough namespaces created with flags to resources
Documentation translated
entry/ Domain documentation resources defined [all]
do: Translate resources to common format. Trace origin
exit/ DOC resources formalised
Core ontologies translated
entry/ Core ontology resources defined [FIGIS, top ASFA, top Agrovoc, else]
do: Translate core resources to common format
exit/ Preliminary core ontology formalised
Topic trees translated
entry/ Topic resources defined [oneFish]
do: Translate resources to common format
exit/ Preliminary topic trees formalised
Lexical sets translated
entry/ Lexical resources defined [all]
do: Translate resources to common format. Trace origin
exit/ Lexicalization resources formalised
Domain ontologies translated
entry/ Domain ontology resources defined [FIGIS]
do: Translate resources to common format
exit/ Domain ontology resources formalised
BT/NT hierarchies translated
entry/ Domain BT/NT resources defined [ASFA,Agrovoc]
do: Translate resources to common format
exit/ BT/NT resources formalised
RT axioms translated
entry/ Domain RT resources defined [ASFA,Agrovoc]
do: Translate resources to common format
exit/ RT resources formalised
RT resources refined
entry/ RT resources formalised
do: Refine with heuristics based on taxonomies and core ontologies
exit/ Refined subset of RT axioms ready
BT/NT hierarchies refined
entry/ BT/NT resources formalised
do: Refine with heuristics based on core ontologies
exit/ Refined subset of BT/NT hierarchies ready
Topic trees refined
entry/ Preliminary topic trees formalised
do: Refine trees according to ontoclean principles
exit/ Refined topic trees ready
Assertional resources ready
entry/ Domain ontologies translated, BT/NT hierarchies refined, RT resources refined, DOC and lexicalization resources formalised
do: Prepare integration space of assertions
exit/ Assertions to be integrated
Axiomatic resources ready
entry/ Domain ontologies translated, RT axioms refined
do: Prepare integration space of axioms
exit/ Axioms to be integrated
Taxonomical resources ready
entry/ Domain ontologies translated, BT/NT hierarchies refined
do: Prepare integration space of taxonomies
exit/ Taxonomies to be integrated
List of integratable ontology elements ready
entry/ Taxonomical, axiomatic, and assertional resources ready, refined topic trees ready
do: Create working namespaces with flags to original resources, maintain links between current resources
exit/ Working, interlinked namespaces created with flags to resources

Fig. 2. A state diagram that sketches the methodology used to extract and refine the informal data.

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'.

3.2 Translation and refining of components for IFO building

The 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] would 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* (R-CO.1).
- Translate FIGIS reference tables: taxonomy, individuals, and local relations (to be transformed into formal axioms). This results in a *resource for domain ontology development* (R-DO.1).
- 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* (R-OL).
- Extract IS_A taxonomies from AGROVOC and ASFA BT/NT (*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.* This results in *resources for core and domain ontology development* (R-CO.2,3, R-DO.2,3).
- Expand RT (*Related Term*) relations from AGROVOC and ASFA (heuristics from IS_A taxonomies is to be used). Also non-IS_A BT/NT hierarchies could be refined (expanded) here. This results in *resources for core and domain ontology development* (R-CO.4,5, R-DO.4,5).
- Reuse existing documentation: oneFish topic summaries, AGROVOC and ASFA scope notes, FIGIS glossary. Consider that documentation can be used at development time (axiomatisation, cf. §4.3.2), as well as at runtime (informal description). Runtime documentation needs a versioning tool to maintain consistency with source glossaries. Specialised ontological documentation should be provided, specially for core ontologies. This results in *resources for ontology documentation* (R-GL.1,2,3,4).

- 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 <u>*resources for ontology lexicalisation*</u> (R-LEX.1,2,3,4).

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)

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. Resources:

Upper ontology resources: the OntoClean upper level [8] 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, (e.g. R-CO.1,2,3,4,5), relational database schemas, theoretical views on domain topics, domain standards, etc.

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*

Development is performed as incremental loading and classification of upper and core level ontologies in the Ontology Server. This results into the secondary resource SR-FCO.

3.3.2 Building domain IS-A taxonomies.

Integrate the resources for domain ontology development (R-DO.1,2,3) with the fishery core ontology (SR-FCO).

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 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 the secondary resource SR-DTA (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 SR-RDTA (Refined DTA).

3.3.3 Collect existing documentation and produce glosses.

Integrate the resources for ontology documentation (R-GL.1,2,3,4).

For concepts lacking a gloss, produce a new one.

For core concepts and relations, besides existing glosses, an extensive description of their scope in the FCO should be provided. This results into the secondary resource SR-GL.

3.3.4 Designing a preliminary topic architecture.

Figure out a preliminary topology for most general topics (to be used for ontology modularisation as well). Resources:

Ontologies for topics (Welty's topic topology [15], topic maps standard [16], OnTopic principles [11], semantic portals design [17]).

oneFish topic trees (R-OL).

This results into secondary resource SR-OL.

3.4 Building domain axioms

Once taxonomies are cleaned to a certain extent, documented, and divided into appropriate namespaces, some activities aimed at raising the conceptual detail of the ontology can be started. The most important is the characterisation of domain concepts with axioms.

3.4.1 Integrating resources R-DO.4,5 and upgrading them to the status of logical axioms (formalise informal axioms).

This requires understanding the quantification applicable to those axioms: existential (necessary) or universal (contingent)?

This results into secondary resource SR-DAX.1 (upgraded Domain Axioms).

3.4.2 Axiomatising glosses from SR-GL.1,2,3,4.

Here the ONIONS methodology [10] can be applied to derive formal domain axioms from natural langage descriptions.

This results into secondary resource SR-DAX.2.

Warning: this activity is time-consuming, and semi-automatic techniques are still a research issue [13]. Scalability and approximate results should be considered for the final project phases.

3.4.3 Revising and harmonising formal descriptions from SR-DAX.1,2 according to conceptual schemata (FCO).

This results into secondary resource SR-DAX.3.

3.5 Modularising ontology library according to topics

Reconstruct dependency chains in SR-DAX.3 and check preliminary topic topology (SR-OL) to produce a first version of the ontology library architecture (OLA). Here the OnTopic methodology [11] can also be applied to derive boolean search spaces from dependency chains of topics.

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 these 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.

R-LEX.1,2,3,4 are integrated according to SR-RDTA.

3.7 A mediation architecture

Figures 3 and 4 show two simple architectures to support information brokering [6] or unified search after merging of fishery information systems by means of Fishery Ontology Service.

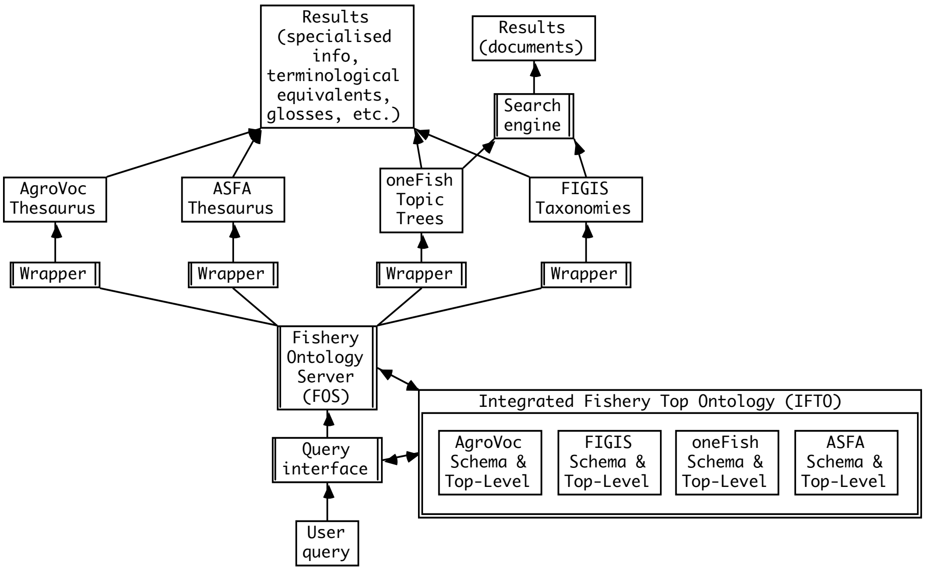

Fig. 3. A brokering architecture for querying heterogeneous fishery ISs.

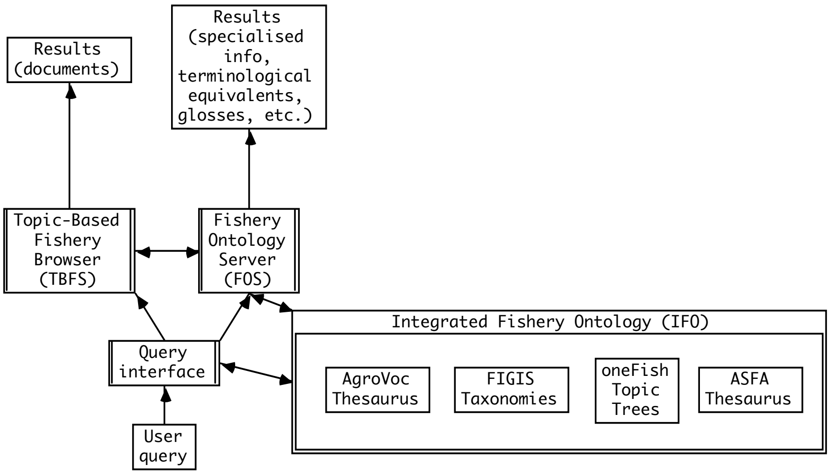
Results (documents)
Results (specialised info, terminological equivalents, glosses, etc.)
Topic-Based Fishery Browser (TBFS)
Fishery Ontology Server (FOS)
Integrated Fishery Ontology (IFO)
AgroVoc Thesaurus
FIGIS Taxonomies
oneFish Topic Trees
ASFA Thesaurus
Query interface
User query

Fig. 4. A unified interface after merging of heterogeneous terminological resources.

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 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. IP-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
17. Stojanovic N, Maedche A, Staab S, Studer R, Sure Y: SEAL – A Framework for Developing SEmantic PortALs