This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

AQUATIC SCIENCE SUBJECT GATEWAY PROJECT AS A MODEL OF INTEROPERABILITY

Carmen Reverté Reverté. Aquatic Ecosystems. IRTA. E-mail: carme.reverte@irta.cat

Dra. Montserrat Sebastià Salat. University of Barcelona. E-mail: msebastia@ub.edu

Abstract

The ASESG project involves design of an interoperable and specialized information system. This system has to solve the main problem of any multilingual and multidisciplinary information system, namely semantic interoperability, focused on simultaneous access to different heterogeneous collections between metadata domains and data mapping.

Keywords: *Aquatic Science, Multilingual Information Systems, Semantic Interoperability,* Subject Gateways, *Thesaurus.*

1. Introduction

Building a specialized *Subject Gateway (SG)* in aquatic science is in response to specific user needs and/or the area of specialization identified in the Aquatic Ecosystems Center (IRTA, Spain), where difficulties are evident in the *Semantic Interoperability (SI)*. This concept is related to the services and information systems growing within the digital environment, which interface with multi-management and heterogenic systems. Therefore, one of the most representative systems in this area is the *Subject Gateways (SGs).* It is identified as an academic and research information structure which unifies research communities characterized by the same information necessities. It is in fact a high quality system. In this project we investigate other representative projects and international subject gateways like RDN, Intute, Renardus, Carmen and Australian gateways.

Semantic Interoperability is the most important part of the Subject Gateways *Information Architecture* (IA) because of the fact that they are composed of multiple organizations, several information systems, different subject areas and a multilingual text. The main objective of *SI* is to solve the heterogeneity problems that exist in this kind of digital environment, to share and recycle information and services in order to obtain quality information for the end user. However, four heterogeneity levels contain potential interoperability types:

1. System level or software incompatibilities.
2. Syntactic level or differences between codes and representations of the programs (algorithms or metadata).
3. Structural level or differences among data models, structures and schemes (mainly algorithms, metadata and mapping).

4. Semantic level or terminological inconsistency and meaning differences.

Although, whichever information digital system is used to achieve the four levels, this study is only focused on *Semantic Interoperability*, the key point of accessing quality information within Libraries and *Subject Gateways*.

2. Semantic Interoperability Background

Nowadays, Semantic Interoperability (SI) operates on several levels, but all of them have in common the information management and accessibility within controlled vocabularies and user interaction.

In the librarianship traditional context, *SI* is used for subject indexing and subject access to search support, providing accuracy in the information retrieval process. The necessity of cross-searching within bibliographic online databases in order to subject access from multilingual and multidisciplinary collections.

The second and more recent context is heterogeneous information resource interoperability and their homogenization. Information resources diverseness is increasingly on the Web and this is a problem to get high information quality. Because of that, there is an *IA* Web based on services. The result is the proliferation of multiple information systems: digital libraries, SG's, data warehouses, metadata interchange services, peer to peer architectures, and several knowledge management systems (semantic web and ontologies) and e-Government services [5]. Moreover, information access through subject browsing becomes known as the key in all digital information systems for users.

Another viewpoint is the activity based on 'information life cycle' management. Information is gleaned from such sources as government institutions (eGovernment services), museums, and business. Following this research line on information management systems and archival processes within a digital environment like a SG, we can identify several levels of *SI* with different 'information life cycle' processes:

Table 1: Semantic Interoperability & Information life cycle

Information life cycle	***Semantic Interoperability* processes**	***SI* levels**
Build an information product	Terminologies and metadata use increase the control among different disciplines.	Level 1
Collection development	Controlled Vocabularies use for digital documents/objects inclusion and evaluation.	Level 1 & 2
Cataloguing (indexing, organization and	Use of metadata, controlled vocabularies and authority control. And developing	Level 2 & 3

knowledge management	common polices for description (MARC21).	
Information integration (mapping processes)	KOS interoperability: interchange protocols (Z39.50). Use of metadata standards and RDF schemes (OWL o SKOS) and controlled vocabularies mapping.	Level 1, 2 & 3
User Interaction (user interfaces)	Personalization services, Alert services, support services, etc.	Level 1, 2 & 3
Information access	*Cross-Indexing, Cross-Browsing* and *Cross-searching* with controlled vocabularies. And services diffusion.	Level 1, 2 & 3
Management, Preservation and Archive	Update, verifying links, metadata and standards formats use. Storage.	Level 1

2.1 Potential *SI* problems and solutions

Technically, the *SI* involves complex tasks in multiple levels and functions of the information systems. However, the process can be identified on three levels (Doerr, 2004):

Data structures, describe states of affairs, information control and management functions are normally local to a system where they are the fields and data tables. This is the metadata, content data, collection management and service description data. From an ontological point of view the data structures are related to universals of the domain but not to particulars, for this reason it is necessary to use metadata standards and interoperability languages. They are a relevant component of the *SI* because it contains information about the resource (title, data, authority, typology, etc.).

First of all, the *SI has to* guarantee the data structures mapping. It is associative processes between data elements and structures. The most known metadata mapping process is *crosswalk:*

Table 2: Crosswalk example: mapping between metadata formats

Marc Fields	Dublin Core Elements	NOAA FGDC
245$a (Title) *245.10 $a Aquatic plant book*	<**DC:title**>*Aquatic plant book*</>	**1.1.8.4 (**Title) *Aquatic plant book*
100,110,111,710,711 (Author) *Ex.: 100.10 Cook, Christopher D.K.*	<**DC:creator**>*Cook, Christopher D.K.* </>	**1.1.8.1 (Originator)** *Cook, Christopher D.K.*

260$a (Publication Place) *Amsterdam*	<**DC:publisher**>*Amsterdam*</>	**1.1.8.8.1 (Publication Place)** *Amsterdam*
260$b (Publisher) *SPB Academic Publishing $c 1996*	<**DC:publisher**>*SPB Academic Publishing*	**1.1.8.8.2 Publisher** *SPB Academic Publishing*
260$c (Date) *1996*	<**DC:data**>*1996*</>	**1.1.8.2 (Publication date)** *1996*
650$a (Subject); 650 2 / 653 (Subject)*; Ex: 650.04 Freshwater plants $x Identification*	<**DC:subject**>*Freshwater Plants*, *Identification*</>	**1.6.1.1 (Theme Keyword)** *Freshwater plants – Identification***;** **1.6.1.2 (T.K. thesaurus / term uncontrolled)**

This mapping process is possible with automatic tools, algorithms that can transform an **A** structure with a **B** or **C** data structure. Moreover the metadata not always include enough subject information, and it is necessary to include factual and categorical data too.

Categorical Data are universal dates, standards dates like controlled vocabularies and terminologies used for systems accuracy. Comparing with 'data structures' their standardization is more difficult and sometimes impossible. Because terminologies are related to user communities (culture, language and ideology) and their structure equivalence not always is possible.

On the other hand, problematic scenery is the Internet growing and subject proliferations like browsing systems. Moreover, catalogs are more specialized and interdisciplinary than before; consequently their indexation is less deep. Difficulties for accessing to several levels of granularity of the multidisciplinary system, and constant evolution of controlled vocabularies are showing heterogenic scenery which we must face with *SI* way:

1. **Common use of classification systems:** documents re-classification/re-indexation using controlled vocabularies or keyword lists which are interoperability systems to information access. It is necessary an integrated access with a mapping process from controlled/ uncontrolled keywords to other keyword systems. Methodologically, it is a really difficult and expensive process, and most of the mapping systems are using classification schemes like a mapping language because is easier mapping numeric systems than terminological systems.
2. **Modeling**: controlled vocabulary specialized or not is developing through another more suitable that it have already existed.

3. **Mapping** (intellectual process): establishing process of terminological equivalences among controlled vocabularies, their structures and relationships. As a result, we have a common suitable system that is able to face the heterogeneity of the multilingual and multidisciplinary systems.
4. **Adaptation/Translation**: controlled vocabularies translation between different languages without changes.
5. **Support** in classification and indexing processes between systems to provide partial interoperability or superposition (not mapping):
    - Metadata enrichment, authorities control and use of classifications and indexes.
    - Semiautomatic classification process to subject access
    - Common indexing systems use
6. **Nothing:** *SI* depends on indexation full text process and indexed references from online systems (bibliographic databases, social nets and Web search engines).

Factual data are *particular* dates. They only can appear once in a digital system. For example, an author relation with a place, date and document. Contrary to categorical data, we can differentiate factual data through codification rules (except for geographical data) to identify repeated items, and through the description of *particulars* with authority standards. It is also important algorithms used for identifying duplicates, *data cleaning* systems and *data mining* techniques as an indexation tool and mapping support.

On the other hand, *SI* has two research lines to bear in mind: *Standardization* and *Interpretation.*

Standardization is a proactive process, everybody can share and access to the data using a common standard. In *SI* the standardization process is related to metadata meaning and schemes for sharing the same KOS concept, authority controls, geographic names and common identifiers. It is the most stable process in a large period of time, but not always is flexible, for this reason is more useful in general information systems.

Interpretation is reactive, it is based on translation, mapping or correlated processes of metadata, content standards (*crosswalk)* and controlled vocabularies like interpretative tools of several information sceneries. It is more flexible and selective and it is used in an environment of high level of diversity (multilingual information systems, SG's and specialized digital libraries). It isn't stable in a large period of time because of changes in research areas.

In the SG's environment the fundamental techniques used are interpretative focused in controlled vocabularies or KOS mapping and *crosswalk*. The

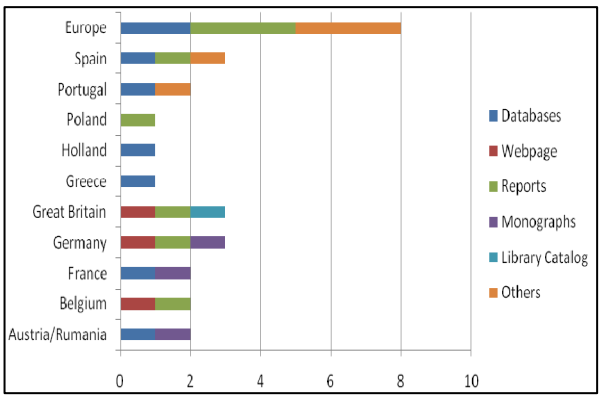
relevancy and accuracy in *Information Retrieval* (IR) are such important as information access, and it is necessary the use, interchange and consistent terminology to obtain a full *SI*. Probably, we need both perspectives (*standardization* and *interpretation*) to assure *SI* in a complex environment (multilingual and interdisciplinary) like ASESG project. On one hand, is necessary to use standard systems for resource description (AACR2, Dublin Core, classification schemes and LCSH), on the other hand we need *interpretation* systems for indexing, browsing and searching between different collections (mapping process).

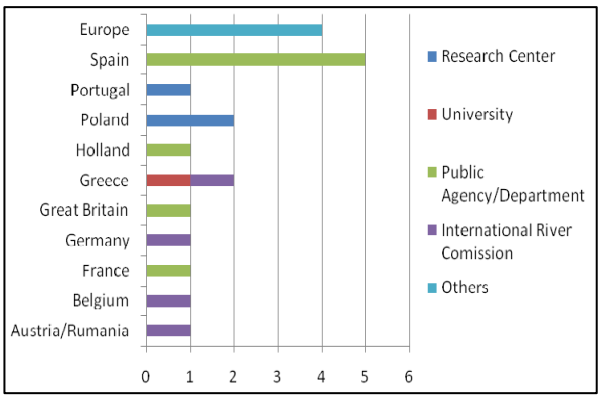
3. The ASESG Project

The objective of the project is developing a framework for a management and *IR* quality system specialized in aquatic science. Aquatic science is a multidisciplinary area, which involves a lot of related topics (Agricultural, Aquaculture, Fisheries, Limnology, Marine Science, Environmental Science and Ecology). Because of this multidisciplinary field, the information resources are found in a heterogenic environment. They are dispersed within several institutions and information systems (research centers, universities, scientific nets and communities, specialized libraries and information centers, databases, portals and others). Furthermore, the fact that we are developing a European project involves designing a multilingual and multidisciplinary information system. Because of that the *Information Architecture* of SG's based on cross-indexing and cross-browsing is the most suitable solution. In this way, the most important point to study is the controlled vocabularies related to the *SI*. In this research line, most of the results obtained raise the hypothesis for building a new aquatic science thesaurus like a controlled vocabulary for ASESG.

For instance, a common situation of information dispersion is delocalization and lack of cooperation and information systems. An information request example about it is "finding data about European rivers water quality", where we can verify that the information management and organization is very different depends on the country. And there is only one case where information is in a library catalog.

Graphic 1: Information resources Heterogeneity **Graphic 2: Institutions Typology**



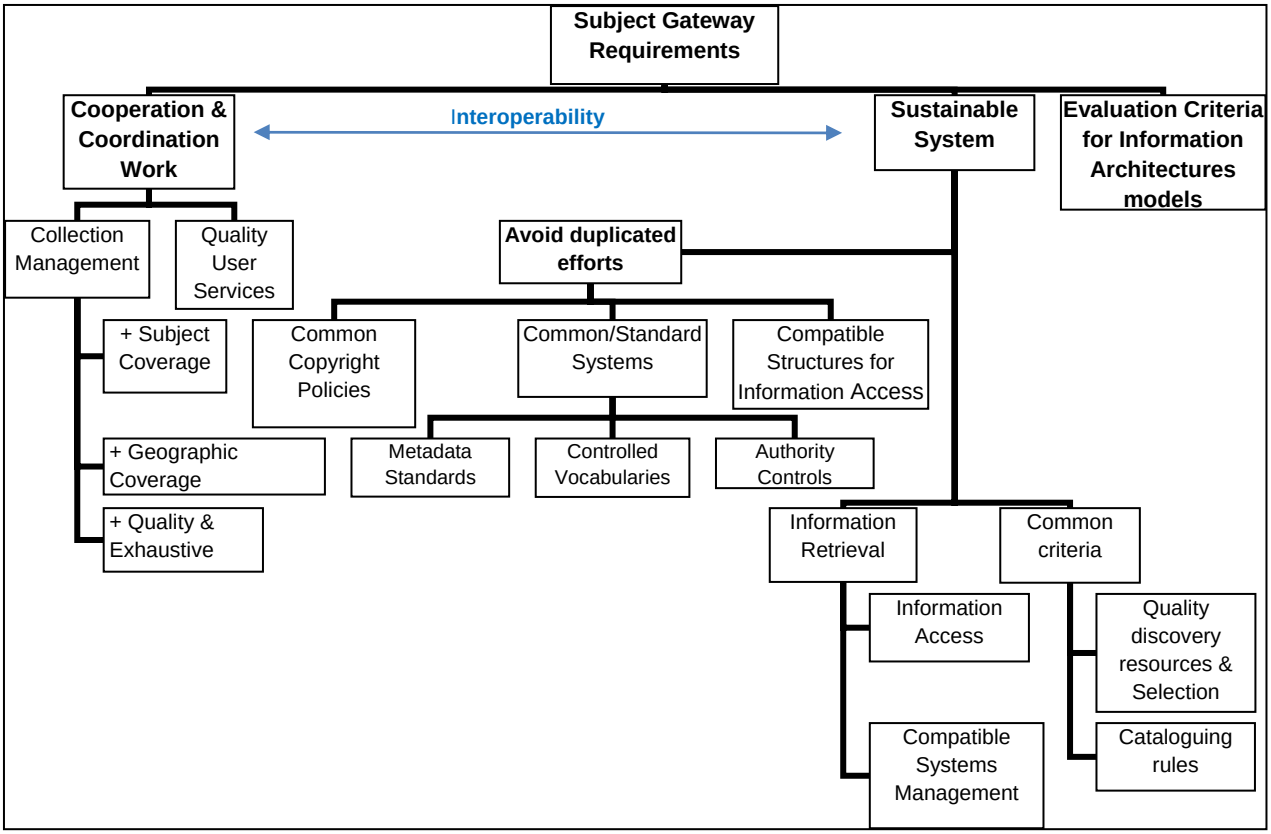


In the ASESG project has developed six stages: analysis *subject context characterization*, *information resources analysis*, *aquatic science controlled vocabularies evaluation*, *SI analysis between the most relevant controlled vocabularies for ASESG*, *study of the most representative SG'S and their AI* (standards, protocols, interoperability, cross-browsing, cross indexing and cross searching systems), and a last stage to build the ASESG framework basis. Despite we are working in the last stages, we have already thought with the potential future partners to develop the project, the aquatic and marine science libraries and information centers nets EURASLIC (Europe) and Medlibs (Mediterranean).

3.1 *Subject Gateway* Information Architecture as a model of *Semantic Interoperability*

Information Architecture mechanisms, which are doing possible that several systems coexist in a *Subject Gateway,* are *cross*-indexing, *cross-browsing* and *cross-searching.* Obviously, they are necessary interoperability protocols, interrogation language, record syntaxes, metadata schemes, controlled vocabularies and cataloguing and description rules between others standards. But the main characteristics that describe the SG's as information quality systems are three: *Cooperation*, Coordination and *Sustainability.*

Figure 1: Subject Gateway essential characteristics for interoperability



3.2 General *SI* requirements in a SG:

Following the most European and International representative *SG's* projects (Australian SG's, RDN project, Intute, Desire, Renardus, Vascoda, etc.), we can find several *SI* levels and requirements:

3.2.1 System level:

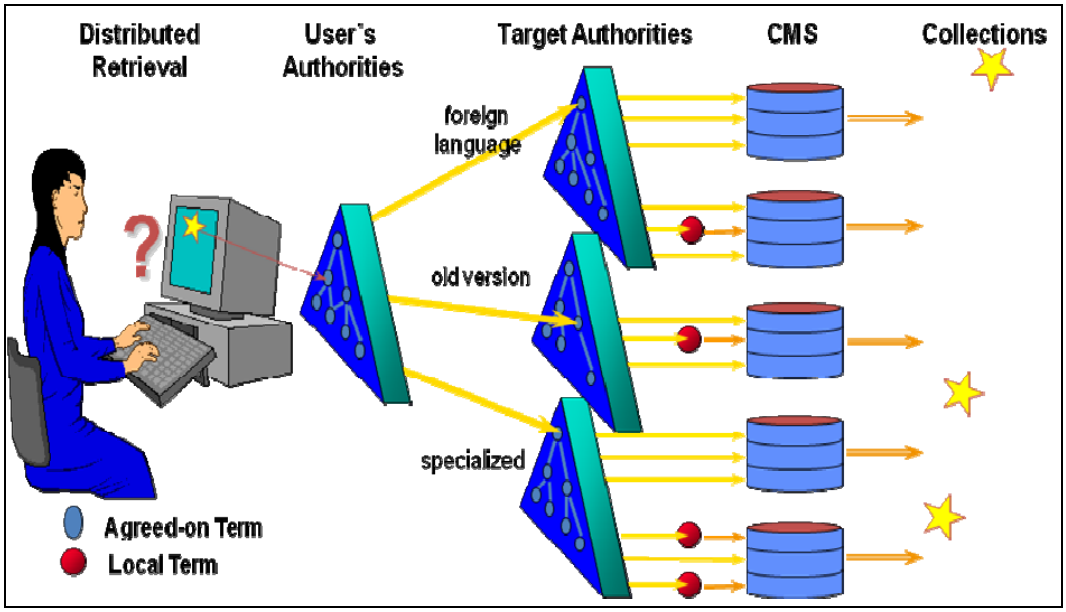
- Protocols and Systems: HTTP, SOAP, Z39-50, OKBC, JDBD, OAI-PMH (Open Archives Initiative Metadata Harvesting) and CIP (Common Indexing Protocol)
- Syntaxes: XML, HTML, Zthes, DTD, SRW i SKOS-Core (RDF scheme)
- Modeling: RDF, OWL (Web Ontology Language), UML
- Semantic: MARC, Dublin Core, IEEE LOM, CIDOC CRM, MPEG-7,

3.2.2 Collection access and interrogation level: *cross-browsing and cross searching*

Cross-searching, jointly with metadata standards (Dublin Core), is used for resources description and language of *RI*. So, we have to evaluate the common KOS schemes involved in the *SG*, know cataloguing rules and tools, and description rules used for each partner.

Cross-indexing process is made when the user address to different digital collection within the same system and using only one controlled vocabulary for *IR and indexing* processes which is mapping to other controlled vocabularies.

Figure 4: Cross-indexing process, *Doerr (2000).*



The objective is able to guarantee the indexing and *RI* consistency among several collections. As a result, they should have a full equivalence among controlled vocabularies (concepts, terms and relationships).

The best situation is that all collections are using common protocols of controlled vocabularies, which can be, manage in their own server. The controlled vocabularies specifications of interchange formats are: authority MARC21, XML (formats based on Zthes DTD) and RDF representations (SWAD-Europe or SKOS-Core).

3.3 Mapping problems:

Multilingualism: the problem is when you have to translate a common language or controlled vocabulary, because there is lack of accuracy, there aren't full equivalences among terms and there is low subject coverage (different levels of coverage or specificity). Moreover, there are polysemy and synonymy problems.

Heterogeneity problems are related to cultural diversity. Vocabularies integration in several languages entails certain risk of conceptual differences.

On the other hand, due to heterogeneity problems with the hierarchical structures of controlled vocabularies, it is difficult to guarantee the equivalence relationships among languages, the thematic depth, accuracy and consistency. Therefore, it is possible to find structural differences among languages (semantic, syntax, lexical and specificity different levels). Moreover, translation uses entails information losses because the conceptual structures of knowledge are different in each language (oriental and accidental world).

Table 3: Heterogeneity Thesaurus Structures & Relationships

ASFA Thesaurus	NBII Thesaurus	GEMET Thesaurus	AGROVOC Thesaurus
Medi Aquàtic **BT Medi Ambient** BT Medi bentònic **BT Medi ambient d'aigües salobres ≠** BT Medi ambient epònic **BT Medi ambient de las aigües continentals** BT Medi intersticial **BT Medi ambient marí**	**Aquatic environment** **BT Environments** **NT Aquatic saline environments** NT Bentic environments NT Compensation depht NT Epontic environments NT Eutrophic environments **NT Inland water environments**	**Aquatic environment** **≠ BT Natural environment**	**Aquatic environment** TR Aquatic communities TR Freshwater ecology BT Environment NT Abyssal environment NT Benthic environment **NT Brackishwater environment** **NT Inland water environment** **NT Marine environment**

4. SG's tendencies in aquatic science

Aquatic Science Subject Gateway hasn't existed yet but we are based on other relevant SG's and aquatic science information centers and libraries environment and their tendencies in *SI* to be able to design a good model. In a technological level we cannot do a depth analysis of their information architecture because

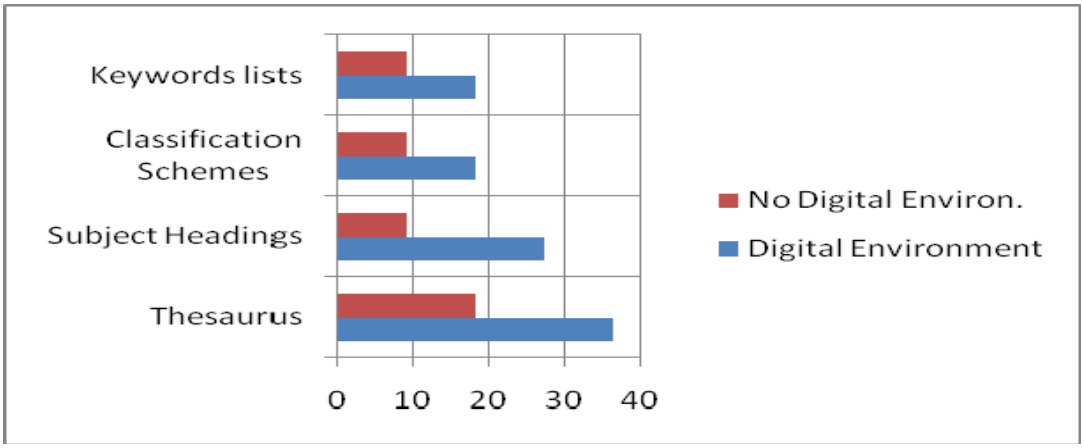
there isn't enough information available. But semantically, both contexts are using more specialized thesaurus for indexing and information retrieval processes than other vocabularies. On the contrary, there are other *SG's, which are using classification systems (Renardus)* or Subject Headings (*INTUTE)* for browsing, indexing and *IR* processes.

Classification systems are used more in multidisciplinary and national context, because the classification schemes are standards and they are more global by definition. Their numeric structure makes easier mapping processes than other controlled vocabularies; on the contrary they have less accuracy.

On the other hand, thesauruses as indexing systems are more complex and more pertinent. They are used in specialized environment (INTUTE, VASCODA, Auslit, etc.). Moreover they can fuse thesauruses, metathesauruses and cross concordances, for this reason they have a great subject coverage and depth.

Regarding to *Aquatic Science* research field, we have done studies within MedLibs (Mediterranean Marine and Aquatic Libraries and Information Centers Network), consist of 45 members. Mainly, in this study we have evaluated the controlled vocabularies diversity, typology and digital or non digital environments uses. The results of this study show us the predominant use of thesauruses as an indexing language.

Graphic 3. Aquatic Science controlled vocabularies typologies



The results show us five representative thesauruses in the aquatic science, but we still have to emphasize the important use of *Subject Headings* (LCSH) and other local vocabularies in information digital systems. Most of times both controlled vocabularies are used together. And one more time, we can verify the heterogeneity in the aquatic science field. In the next table, is showed the

multidisciplinary of aquatic science thesauruses involved in several knowledge areas: *Biology*, *Agriculture, Environment,* Economic and Social Development.

All of them are made with international standards and fulfill a *SI* important requisite, the multilingualism (English as a common language).

Table 4. Aquatic Science thesauruses used in MedLibs

TITLE	INSTITUTION	SUBJECT FIELD	ONTOLOGY (SKOS&RDF)	Languages	Structure (all HIerarchical & Associative
ASFA Thesaurus	FAO (International)	Aquatic Science &Fisheries Abstracts	YES	Multilingual (English, French & Spanish)	**BT, NT, RT, UF, SN**
AGROVOC Thesaurus	FAO(International)	Agricultural thesaurus with Fisheries and Aquaculture part	YES	Multilingual (17 languages)	**BT, NT, RT, UF, SN**
GEMET Thesaurus	EEA (Europe)	Environmental Thesaurus	YES	Multilingual (22 languages)	**BT, NT, RT, UF, SN, Groups and Themes**
NBII Thesaurus	CSA; U.S. Geological Survey's Biological Informatics Office (USA)	Biological, Ecological and Environmental Science	YES	Monolingual-English SOAP Web Services:ASFA, Life Sciences, Pollution and Sociological and CERES/NBII Thesauruses	**BT, NT, RT, UF, SN, SC**
OECD Thesaurus	United Nations (International)	Economic and Social Development	NO	Multilingual	**BT, NT, RT, SN**

4.1 Thesaurus mapping tendencies

Nowadays, controlled vocabularies integration is used like management, achievement, indexing and *IR* tools for their future retrieval in digital sceneries. They are following the ontological research line, overall, in multilingual information systems and heterogenic collections, where controlled vocabularies are converted to data schemes like metadata standards (*Dublin Core)*. Therefore, *ontologies* are languages with higher degree of *SI* among systems, even though such the data interchange as their mapping is faster and easier.

Through a previous study of context and target, we obtained heterogenic data and great split of information resources in aquatic science field. On the contrary, there are a lot of scientific communities and international specialized

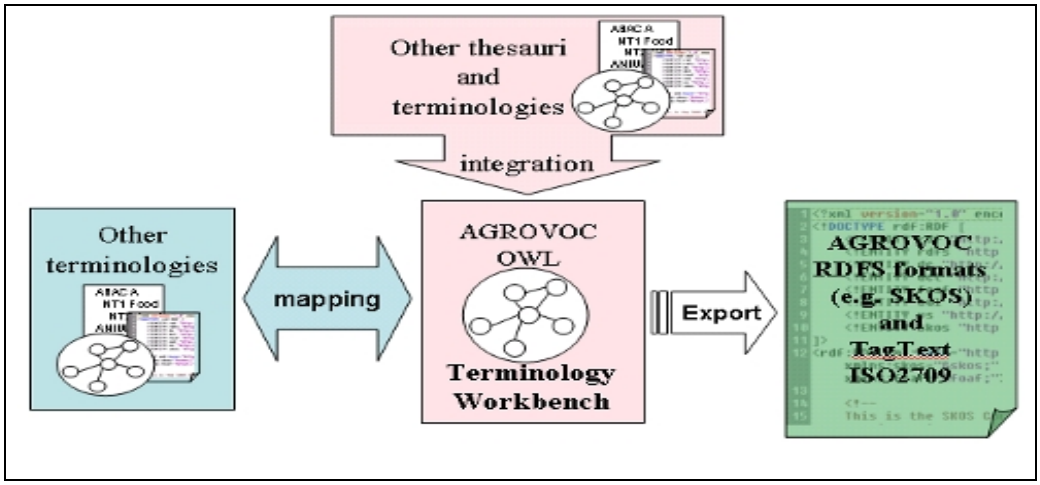
librarianship nets with the same interest and objective to provide information access and diffusion because the lack of information systems.

As for *SI* and controlled vocabularies integration we are following relevant European and International aquatic science projects:

CERES/NBII Thesaurus (California), which is developed a standard protocol for thesaurus distribution and communication through mapping processes of metadata scheme. They built and develop metadata tools, HTTP protocols with RDF thesaurus (XML) and NISO Z39.19 representation standard. Use of Zthes Z39.50 for thesaurus browsing based on ISO 2788. After that Zthes is used for ZING (Z39.50-International: Next Generation within SRW Search/Retrieve Web Service protocol) with SOAP and URL mechanisms.

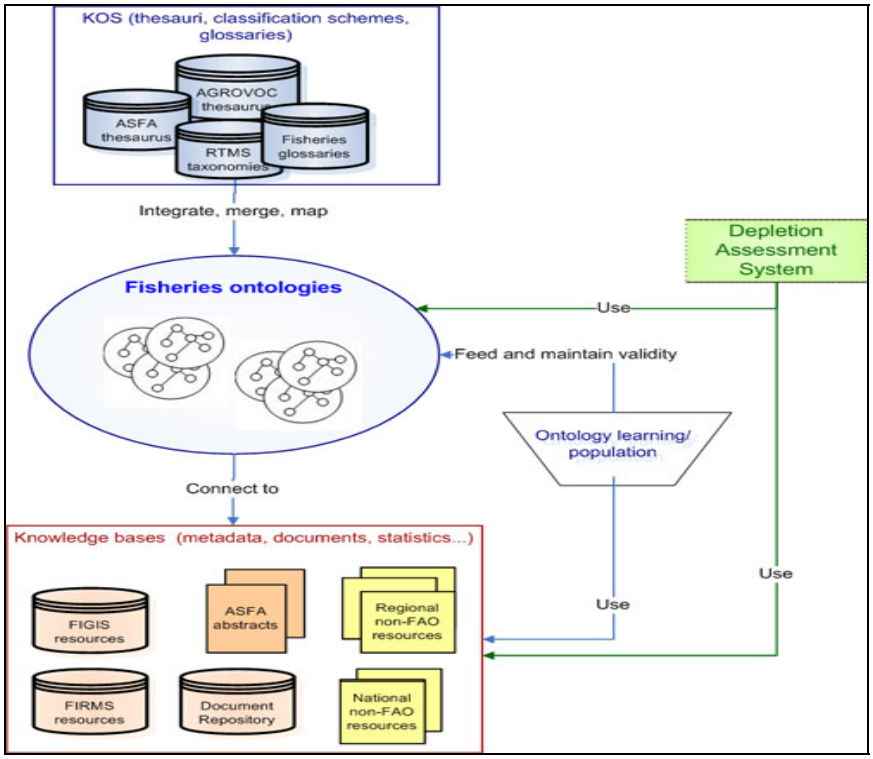
Other important project is AGROVOC (FAO) which converted their thesaurus into ontology through thesaurus data conversion in RDF files with OWL standard format. This process allows to mapping and representing several controlled vocabularies.

Figure 6: AGROVOC mapping process



Other interesting project is Marine Metadata Interoperability (National Science Foundation). It is based on RDF files which are used for the mapping process. This project only includes marine science not aquatic environments (inland or freshwater environments). Last relevant project is NeOn, a marine ontological European project where they are AGROVOC and MMI. In this project there are a mapping project between ASFA and AGROVOC thesaurus. Even though, the mapping processes aren't really optimistic because of subject coverage differences.

Figure 6: *NeOn* mapping process



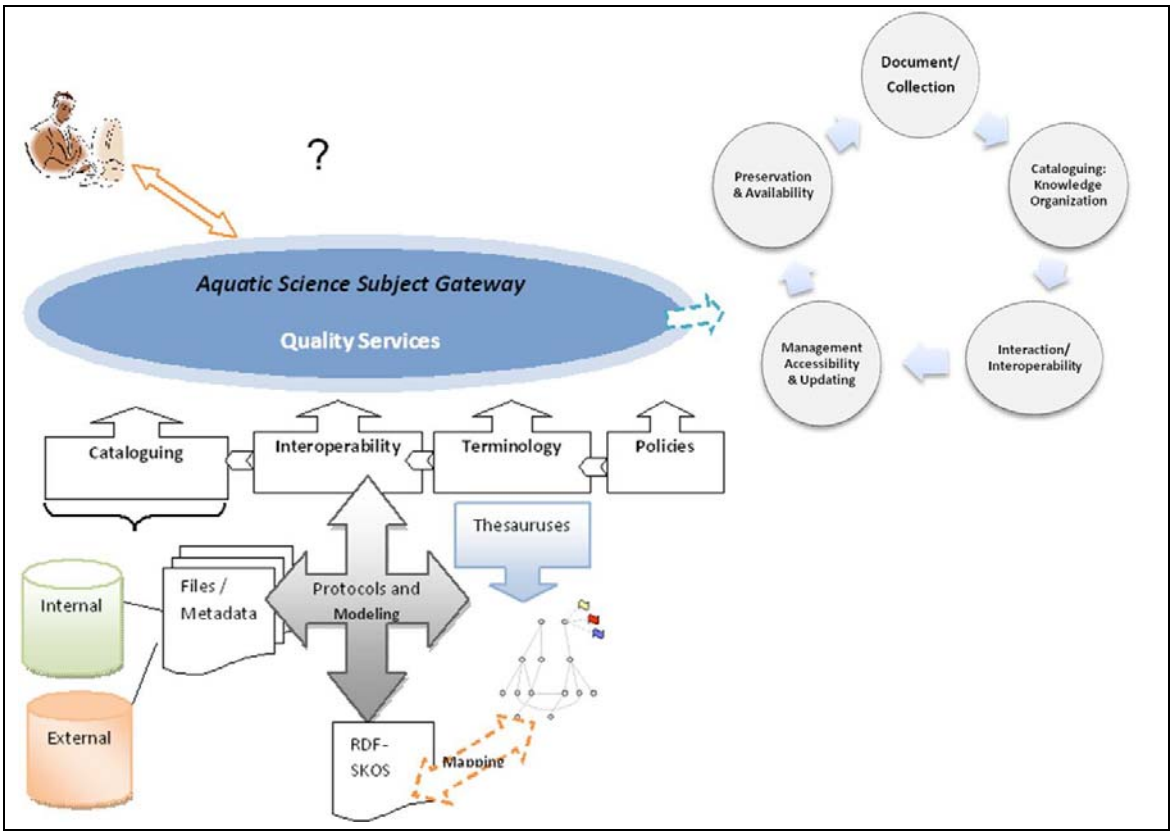
Other research line that we are studying are a new multilingual thesaurus construction on aquatic sciences, a special subject thesaurus for the field of aquatic ecosystems, water and wastewater management. The Aquatic Thesaurus is a collaborative project that involves several European partners' (EURASLIC and MedLibs) which integrates the controlled terminologies of different vocabularies for different domains and tasks. Because of its multilingualism, this thesaurus cannot only be used for indexing, retrieval and terminological reference, it should serves also as a translation tool for the languages represented (at the moment we are considering Catalan, English and Spanish languages). The establishment of cross-concordances related on thesaurus is the basic elements to solve the heterogenic semantic problem (following the FACET project), and it's processing with a Multites Thesaurus Construction program. They enable the treatment of semantic heterogeneity within subject gateways, and the Aquatic Thesaurus reflects the changes in terminology. The *SI* problem is focused on simultaneous access to different collections between metadata domains and data mapping.

5. ASESG Semantic Interoperability Model

One **user** has to search an **information system** which has to interrogate within different **collections** deposited in several data bases which are **mapping** with

interoperability standards criteria's for information contrasting and their later **information retrieval** which is more accuracy and Relevancy.

Figure 7. Semantic Interoperability Model



Moreover, for being a quality system it is necessary realized the 'information life cycle'. So, we have to consider the *SI* from acquisition process or information resource to the final user quality product.

6. Conclusions

ASESG project has outlined some research solutions within framework of *Subject Gateways* and *Semantic Interoperability* integration to cover the lack of aquatic information systems in order to offer quality services for researchers and aquatic science professionals. Some details have been given to solve the heterogenic problems that involve the *SI,* based on several standards (metadata and RDF schemes) and interpretation (*data* mapping) methods.

Overall, this research issue is focused on providing a unified methodology of aquatic science information systems integration based on *Cooperation*, Coordination and *Sustainability* among aquatic science specialized libraries.

7. Bibliography

(2000) 3.6 Interoperability. Desire Handbook. URL: http://www.desire.org/handbook/3-6.html

(2004) Final European Interoperability Framework. European Communities. URL: http://ec.europa.eu/idabc/servlets/Doc?id=19529

(c2002-09) The FACET Project University of Glamorgan. URL: http://reswin1.isd.glam.ac.uk/FACET/live/demo_TermViewer.asp

(c2008) Crosswalks tools in Marine Metadata Interoperability. USA, National Science Foundation. URL: http://marinemetadata.org/tools

DEMPSEY, L., CHILDRESS, E., GODBY, C. G. & AL.], E. (c2004-05) Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape. LITA guide to e-scholarship (working title). url: http://www.oclc.org/research/publications/archive/2004/ dempsey-mslitaguide.pdf

DOERR, M. (2004) Semantic interoperability : Theoretical Considerations. ICS-FORTH. URL: http://www.ics.forth.gr/ftp/tech-reports/2004/2004.TR345_Semantic_Interoperability_Theoretical_Considerations.pdf

KOCH, T. (2006) Electronic thesis and disertations services : semantic interoperability, subject acess, multilinguality. E-Thesis Workshop. Amsterdam.

LAUSER, B., JOHANNSEN, G. & CARACCIOLO, C. (2008) Comparing Human and Autmatic Thesaurus Mapping Approaches in the Agricultural Domain. Proceedings International Conference On Dublin Core and Metadata Application Berlin.

NEUROTH, H. & KOCH, T. (2001) Cross-browsing and cross-searching in a distributed network of subject gateways: Architecture, data model, and classification. Renardus Project. URL: http://old.stk.cz/elag2001/Papers/HeikeNeuroth/HeikeNeuroth.html#chap2

PATEL, M., KOCH, T. & DOERR, M. (2004) Semantic Interoperability in Digital Library Systems. European Comission. URL: http://ec.europa.eu/idabc/servlets/Doc?id=19529