AQUATIC SCIENCE EUROPEAN SUBJECT GATEWAY PROJECT (ASESG) AS A MODEL OF INTEROPERABILITY

Carmen Reverté Reverté. Aquatic Ecosystems Documentation Center. IRTA. (Sant Carles de la Ràpita, Spain). E-mail: carme.reverte@irta.cat

Montserrat Sebastià Salat. Universitat de Barcelona (Spain). E-mail: msebastia@ub.edu

0

1. INTRODUCTION

The **AESG** project involves designing an interoperable And specialized information system. It has to solve the main problem of any multilingual and multidisciplinary information system: the semantic interoperability, Focused on simultaneous access to different Heterogeneous collection between metadata domains And Data mapping.

The **Four** main heterogeneity levels and interoperability problems are:

- System level or software incompatibilities
- Syntactic level or differences between codes and representations of the programs (algorithms and metadata).
- Structural level or differences among data models, structures and schemes.
- Semantic level or terminological inconsistency and meaning differences

"Semantic Interoperability is the main factor to solve the heterogeneity problems in a multilingual and multidisciplinary digital environment"

2. Semantic Interoperability (SI) Background

SI has several stages but all of them have in common "the information management and accessibility" within controlled vocabularies and user interaction.

1.Traditional (Librarianship) context: SI is Used for subject indexing and subject access to **search support** and providing **accuracy** in the information retrieval (**IR**).

2. Second and more recent **context** is **the heterogeneity**. Information resources diversity is increasingly on the Web:

+ Web services
+ Information systems
+ Information access through subject browsing
- Information Quality

3. Third context, followed by museums, eGovernment services and business, is based on 'information life cycle'. Each cycle process is identified with different levels of SI:
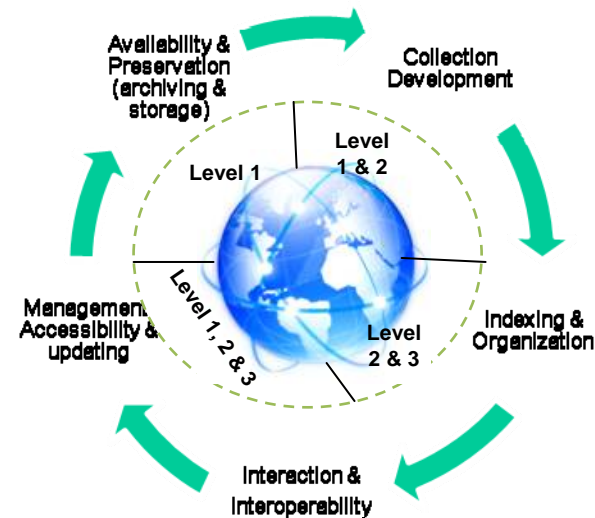

Table 2. Semantic Interoperability Levels (Doerr, 2004)	
Level 1	Data Structures = Metadata
Level 2	Categorical Data = Standards (Controlled Vocabularies)
Level 3	Factual Data = Particular data (Geographical & Authority control)

2.1 Potential SI problems and solutions: 3 levels of SI

1. Data Structures: information control and management functions (metadata, content data, collection management and service description data).

SI solution: data structures mapping through associative processes between data elements and structures (crosswalk):

Table 3: Semantic Interoperability & Information life cycle

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)**

2. Categorical Data: universal dates standards used for system accuracy (controlled vocabularies).

Problems (heterogeneity)	Solutions (according to each institution)
Terminologies (semantic and meaning differences)	- **Common use of classification systems:** mapping process from controlled/uncontrolled keywords to other keyword systems (more difficult and expensive process).
Subject proliferations on the net (browsing systems)	- **Modeling:** one controlled vocabulary is developing through another more suitable that it have already existed.
Catalogues: + specialized & + interdisciplinary than before; consequently their indexation is less deep	- **Mapping** (intellectual process): process of terminological equivalences among controlled vocabularies, their structures, relationships and terminology (a more common suitable system)
Access problems to several levels of granularity of the multidisciplinary system	- **Adaptation/Translation:** controlled vocabularies translation between different languages without changes.
Constant evolution of Controlled Vocabularies	- **Support** in classification and indexing processes between systems to provide partial interoperability or superposition (not mapping).

3. Factual Data: particular dates, they only appear once in a digital system.

Two research lines in SI	Process	Advantages/Disadvantages	
Standardization – Proactive (everybody can share and access to the data using a common standard)	Metadata meaning and schemes for sharing the same KOS: authority controls, geographic names and common identifiers	+ Stable process in a large period of time.	- Flexible. Used in general information systems.
Interpretation – Reactive (interchange and consistent terminology to obtain a full SI)	Translation, mapping or correlated processes of metadata, content standards (*crosswalk*) and controlled vocabularies mapping like interpretative tools within system.	+ Flexible and selective. Used in multilingual and specialized digital information systems.	- Stable in a large period of time because of changes in research areas.

3. The ASESG PROJECT

The Objective: Develop a framework for a management and **I**nformation **R**etrieval (IR) quality system specialized in aquatic science.

Subject Coverage: aquatic science is a multidisciplinary area, which involves several related topics (Agricultural, Limnology, Marine Science, Environmental Science).

Context: heterogenic field which has information resources dispersed within several institutions and information systems (research centers, universities, scientific nets and communities, specialized libraries and information centers, databases, portals and others).

Geographical Coverage: *ASESG* is thought in a European context and it involves designing a multilingual and multidisciplinary information system.

Languages: at the moment only English, Spanish and Catalan.

Proposal: Build a Subject Gateway as the most suitable model to enure the Semantic Interoperability in an information system composed of multiple organizations.

First results: results obtained support the need for building a new aquatic science thesaurus.

The **ASESG** project has developed in six stages**:**

Analysis of the Subject Context Characterization → Information Resources Analysis → Controlled Vocabularies Evaluation → *SI* analysis among Controlled Vocabularies → Study of the representative SG's → ASESG Framework basis
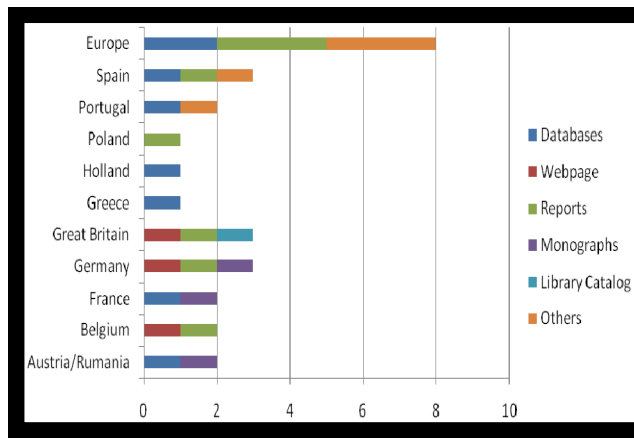
3. The ASESG Project: information dispersion

A common situation is information delocalization and lack of cooperation among organizations and information systems.

An information request example is 'finding data about European rivers quality'. In this request we verified that the information management and organization differs among countries and within the same country too. And there was only one case where information is in a library catalogue.

Graphic 1: Information resources Heterogeneity   **Graphic 2:** Institutions Typology
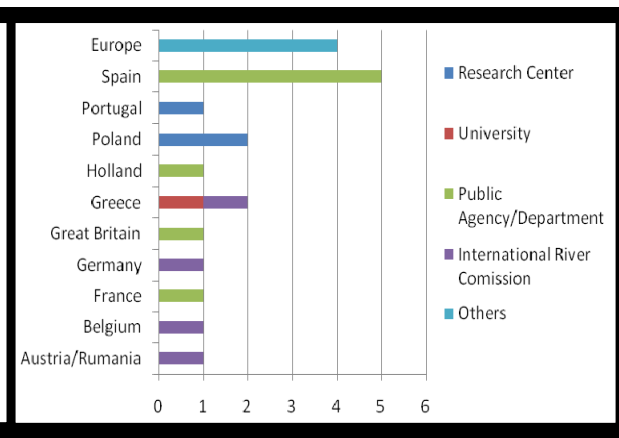
3.1 Subject Gateway Information Architecture as a model of Semantic Interoperability

Cross-indexing, Cross-browsing and Cross-searching are the mechanisms that make possible to coexist several information systems in the same **Subject Gateway**.

The main characteristics that describe the SG's as information quality systems are three: Cooperation, Coordination and Sustainability.
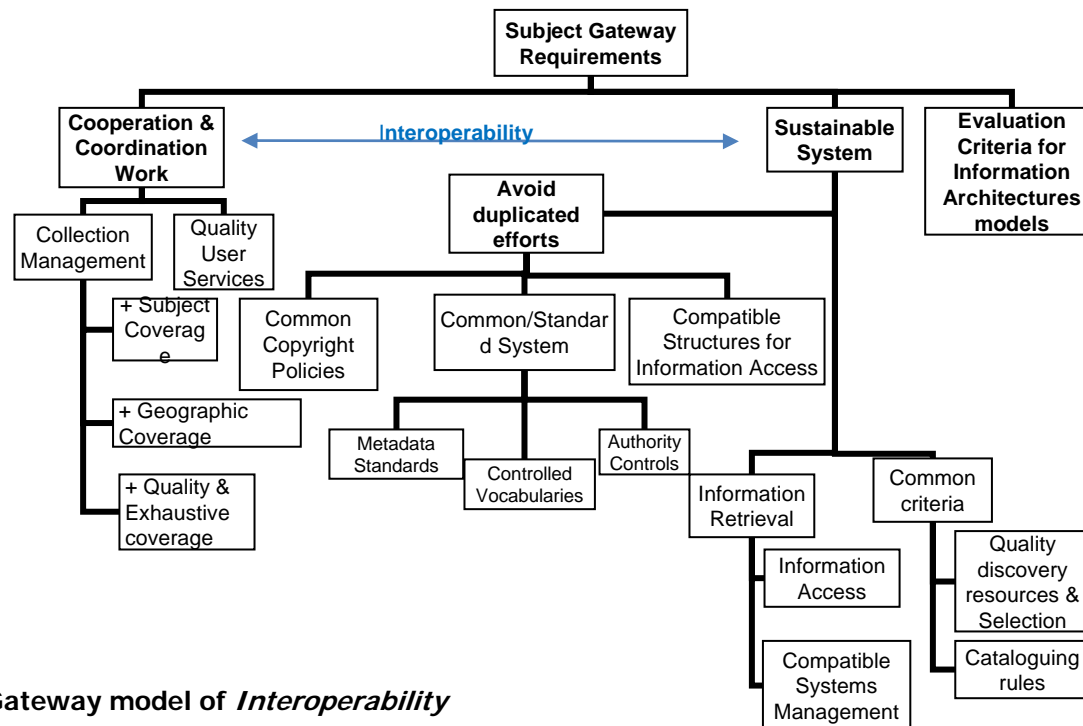

Figure 1. Subject Gateway model of *Interoperability*

3.2 General SI Requirements in a Subject Gateways

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

✓ System level (Syntactic and Structural level)

✓Semantic level (Collection access Interrogation level)

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 and SKOS-Core (RDF scheme)

Modeling: RDF, OWL (Web Ontology Language) and UML

Semantic: MARC, Dublin Core, IEEE LOM, CIDOC CRM and MPEG-7

3.2 General SI Requirements in a Subject Gateways

3.2.2 Collection access and interrogation level: Cross-searching, Cross-browsing and Cross-indexing

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

Cross-indexing and cross-browsing are made when the user addresses to different digital collections within the same system and using only one controlled vocabulary for IR and indexing, which is mapping and other controlled vocabularies.

Figure 2. Cross-indexing process, *Doerr (2000)*
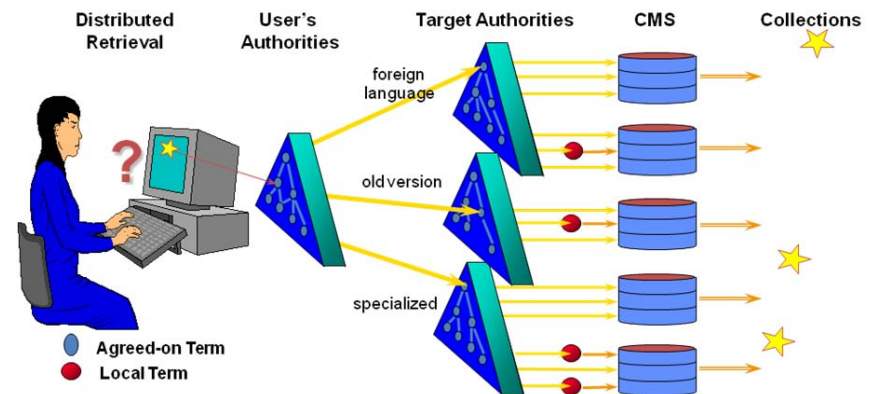
The **Objective** is to guarantee the indexing and *Information Retrieval* **consistency between several collections.**

As a result, they should have a **full equivalence** among controlled vocabularies (concepts, terms and relationships).

3.3 Mapping Problems in SI

Multilingualism problems: a translated common language or controlled vocabulary has lack of accuracy:

• There aren't full equivalences among terms

• Low subject coverage (different levels of coverage or specificity)

• Polysemy and synonymy problems

• Information losses: conceptual structures of knowledge are different in each language (oriental and occidental world)

Heterogeneity problems:

• Cultural diversity: Vocabularies integration in several languages entails risk of conceptual differences

• Hierarchical structures of controlled vocabularies: difficult to guarantee the equivalence relationships among languages, thematic depth, accuracy and consistency

• Structural differences among languages (semantic, syntax, lexical and specificity different levels)

Table 4: Heterogeneity Thesaurus Structures & Relationships

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

4. Subject Gateways tendencies in Aquatic Science

The Aquatic Science Subject Gateway project is based on other relevant SG's and aquatic science information centers and libraries to design a good and efficient *SI* model.
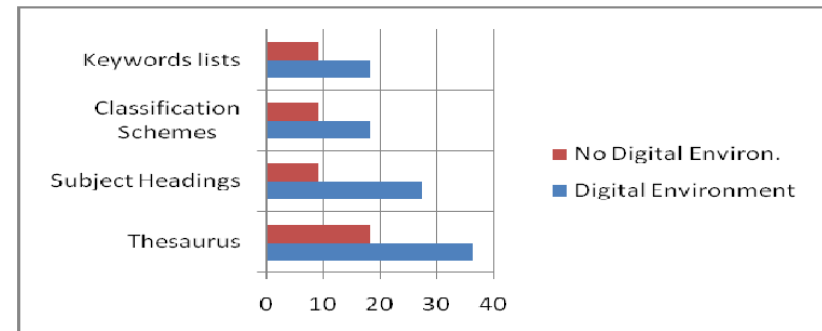
Semantic tendencies in aquatic science libraries are using more specialized **thesaurus** for indexing and **information retrieval processes** than other vocabularies.

Subject Gateways in other areas are using **classification systems** *(Renardus)* or **Subject Headings** (*INTUTE*) for **browsing, indexing and *IR* processes.**

In firsts stages of the project we evaluated the controlled vocabularies diversity, typology and digital or non digital environment uses. Those studies made within MedLibs (Mediterranean Marine and Aquatic Libraries and Information Centers Network).

The results of this study show us the predominant use of thesauruses as an indexing language.

Graphic 3. Aquatic Science controlled vocabularies typologies

4. Subject Gateways tendencies in Aquatic Science

Five representative thesaurus in aquatic science field represent the multilingual and multidisciplinary characteristics of this area.

Table 5. Aquatic Science thesaurus used in MedLib

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

Knowledge areas included: Biology, Agriculture, Fisheries, Aquatic, Environment, Economic and Social Development
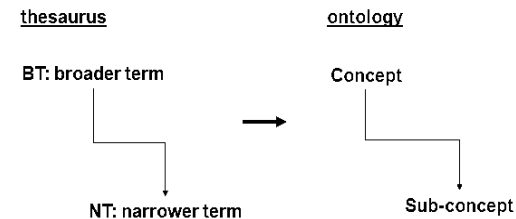
Semantic Interoperability among the thesaurus: international standards, multilingual (English as a common language) and available with RDF format (SKOS language)

4.1 Thesaurus mapping tendencies

Controlled vocabularies integration: are used like indexing and IR tools in digital sceneries following the ontological research line.

Ontology process: controlled vocabularies are converted to data schemes like metadata standards (Dublin Core).

Example:

thesaurus   ontology

BT: broader term   Concept

⟶

NT: narrower term   Sub-concept

4.1 Thesaurus mapping tendencies in Aquatic Science

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

CERES∕NBII Thesaurus (California, 2003): is composed by 6 thesaurus (CSA and CERES thesaurus). Mapping processes through metadata scheme: **HTTP protocols with RDF thesaurus (XML).**

AGROVOC Thesaurus (FAO, 2003): It is converted into ontology through thesaurus data conversion in RDF files with OWL standard format. They are exported to SKOS format (ISO 2709).

Figure 6: AGROVOC mapping process
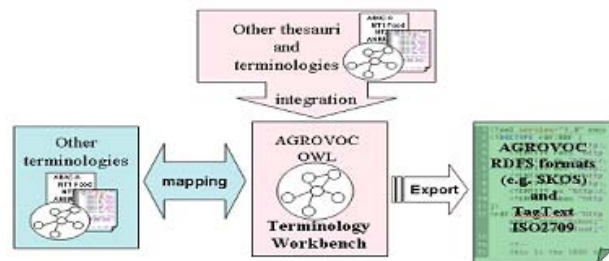
Marine Metadata Interoperability Project (MMI, 2008). Supports collaborative research in the marine science It is based on RDF files for metadata mapping process.

NeOn Project (FAO, 2006): marine ontological European project. One of their mapping process is produced between ASFA and AGROVOC thesaurus. Even though, the mapping processes aren't really optimistic because of subject coverage differences.

Figure 7: NeOn mapping process
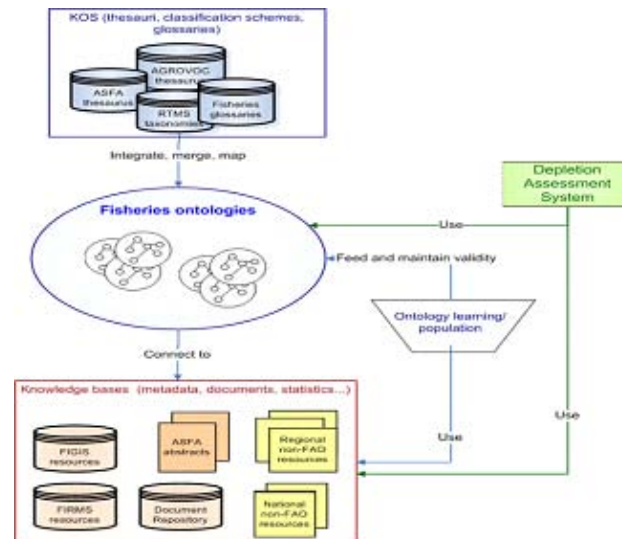
# 4.1 Thesaurus mapping tendencies in Aquatic Science

Preliminary studies and relevant aquatic mapping projects show that coverage of the Aquatic sciences is not enough or sometimes incompatible (NeOn project example)

Establishment of cross-concordances among controlled vocabularies and terminologies will be the basic element to solve the heterogenic semanticproblem. Although, it will be carried out by means of a Thesaurus Construction program (MultiTes) can reflect the changes in Terminology and integrate several controlled vocabularies different from thesauruses (LCSH). A new Aquatic Thesaurus has to be a collaborative project involved of several European partners' (for example, EURASLIC and MedLibs nets)

5. ASESG Semantic Interoperability Model

The user has to search an information system which has to interrogate different collections deposited in several data bases. Data bases are mapped with interoperability standards criteria's for contrasting of information and their later information retrieval, which is more accurate and Relevant.

Figure 9. Semantic Interoperability Model
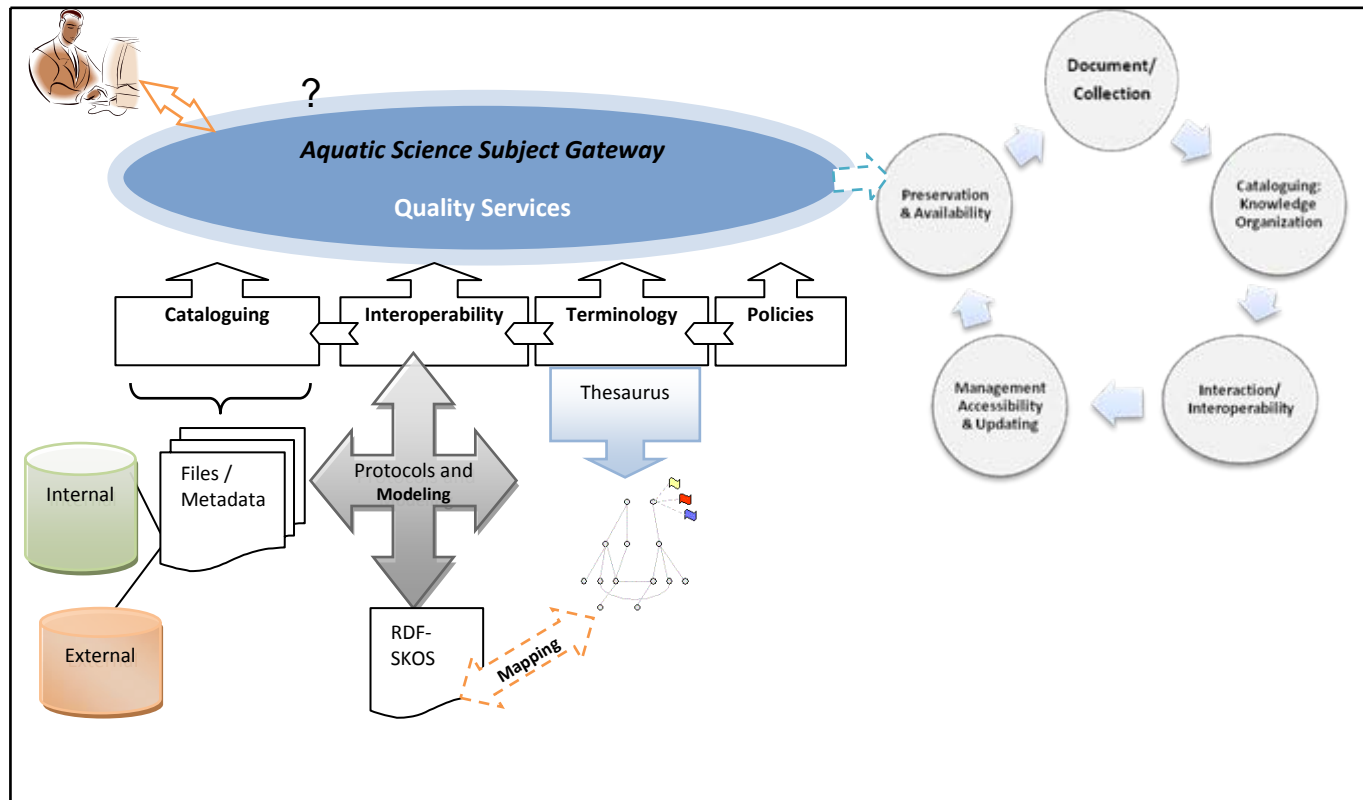
Conclusions

➢**ASESG** project wants to cover the lack of aquatic information systems in order to offer quality services for researchers and aquatic science professionals.

➢To solve the **heterogenic problems** that involve the Semantic Interoperability, we need standardization methods (metadata, RDF schemes and controlled vocabularies) and interpretation methods (data mapping).

➢ **Ontologies**, nowadays, are the best semantic representation and machine understandable representation of knowledge. Could are they the successors of thesaurus and other controlled vocabularies, particularly for information retrieval and knowledge management?.

➢**To assure** the SI in a collaborative environment, we need unify methodologies among aquatic science information systems. And make an integration policy based on Cooperation, Coordination and Sustainability among aquatic science specialized libraries.

THANK YOU FOR YOUR ATTENTION!