Domain Ontology to Support Open Data Analytics for Aquaculture

Pedro Oliveira, Ruben Costa, José Lima, João Sarraipa and Ricardo Jardim-Gonçalves
CTS, UNINOVA, Dep.º de Eng.ª Electrotécnica, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Keywords: Ontology Engineering, Knowledge Representation, Multilingual, Education and Training.

Abstract: The Aquaculture industry, which comprises mainly of SME companies, represents a significant source of protein for people. From an IT perspective, aquaculture is characterized high volumes of heterogeneous data, and also lack of interoperability intra and inter-organisations. Each organization uses different data representations, using its native languages and legacy classification systems to manage and organize information. The lack of semantic interoperability that exists can be minimized, if innovative semantic techniques for representing, indexing and searching sources of non-structured information are applied. The work presented here, describes the achievements under AQUASMART EU project, which aims to accelerate innovation in Europe's aquaculture through technology transfer for the deployment of an open data solution through multilingual data collection and analytics solutions and services, turning the large volumes of heterogeneous aquaculture data that is distributed across the value chain, into an open cloud of semantically interoperable data assets and knowledge. Results achieved so far do not address the final conclusions of the project but form the basis for the formalization of the AQUASMART semantic referential.

1 INTRODUCTION

The aquaculture industry, which comprises mainly of SME's companies, represents a significant source of protein for people. Globally, nearly half the fish consumed by humans is produced by fish farms. Aquaculture is now fully comparable to capture fisheries when measured by volume of output on global scale. The contribution from aquaculture to the world total fish production of capture and aquaculture in 2012 reached 42.2 percent, up from 25.7 percent in 2000 (FAO Fisheries and Aquaculture Department, 2014). Global production is forecasted to increase from 45 million tons in 2014 to 85 million by 2030, making the aquaculture industry the fastest growing animal food producing sector in the world. The European Union needs an innovative aquaculture industry to meet rising seafood demand and to enhance its commercial stocks.

According to the Food and Agriculture Organization of the United Nations (FAO), the volume and value data in global aquaculture production are primarily official statistics obtained directly from the nations and mainly described in local language. Also, there are available other relevant sources of data, like academic reviews, consultant reports and other specialist literature.

The main problems of the Aquaculture sector are related with the lack of global knowledge access, and the inefficient data exchanges and data reuse between aquaculture companies and its related stakeholders. This is primarily due to incompatibility problems among the several information representation structures used by the different software applications along supply chains and business networks (Ray and Jones, 2006). Aquaculture companies have limited capabilities to hire specialized technical resources out of their core business. The main issue fish farmers' face is data understanding and identifying correlations between parameters that affect production, the lack of skilled professionals and the right IT tools, prevents fish farmers to get better insights of their own data and also prevents them to share best practices with other aquaculture stakeholders. For example, if one could reach other growers data (e.g., growth rates, FCR (Feed Conversion Ratio) – related to environmental conditions of cause), than it would be able to have a better and closer prediction plus better and closer feeding according to the current biomass. So, an important step is to be able to get actionable insights in the data resulting in smarter decisions and

Oliveira, P., Costa, R., Lima, J., Sarraipa, J. and Jardim-Gonçalves, R..
Domain Ontology to Support Open Data Analytics for Aquaculture.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 344-351
ISBN: 978-989-758-158-8

better business outcomes, being able to look at past performance and understand that performance by mining the related data (production, environment, etc.) to look for the reasons behind past success or failure and take better decisions for the future.

Knowledge transfer goals are to take the state of the art in multilingual data collection tools, analytics solutions and services, semantic interoperability methods and data mining procedures to implement a global open aquaculture data information system, able to respond to specific user needs and profiles.

Users in the Aquaculture industry can innovate taking the novel capabilities for seamless and holistic access of multilingual data products and services in the Aquaculture value chain, bridging across borders, languages, industries and sectors, removing barriers both technical and organizational.

The access of this global knowledge creates significant new and further commercial opportunities and positive economic prospects.

In India several ICT initiatives are being tried for disseminating aquaculture information to fish farmers, shortening the digital gap and helping the farmer in reaping a good harvest.

There is a demand for intelligent world-class solutions capable of reinforcing partnerships and collaborations with an improved cross-cultural understanding. However due to the proliferation of terminology, organizations from similar business environments have trouble cooperating, and are experiencing difficulties exchanging electronically vital information. To address these issues it is essential to develop semantic tools that easy these integration processes.

Some examples of closely related research streams in recent years are: the extensive work on knowledge models and knowledge management tools, the rise of so-called knowledge engineering, the myriad of projects around 'controlled vocabularies' (such as ontologies, taxonomies, dictionaries, and thesauri), and the academic knowledge-centred courses (graduation, master, and doctoral).

A controlled vocabulary is a list of terms that have been enumerated explicitly. This list is controlled by and is available from a controlled vocabulary registration authority. All terms in a controlled vocabulary should have an unambiguous, non-redundant definition. This is a design goal that may not be true in practice. It depends on how strict the controlled vocabulary registration authority is regarding registration of terms into a controlled vocabulary. At a minimum, the following two rules should be enforced (Lima et al., 2007):

- If the same term is commonly used to mean different concepts in different contexts, then its name is explicitly qualified to resolve this ambiguity.
- If multiple terms are used to mean the same thing, one of the terms is identified as the preferred term in the controlled vocabulary and the other terms are listed as synonyms or aliases.

In line in the notion of controlled vocabularies, an ontology for the Aquaculture domain is being developed. The approach presented here will be beneficial for aquaculture companies in several ways: (i) by enabling to benchmark similar companies with relation to their production performance indicators; (ii) providing access to consolidated information on best practices and success stories; (iii) to be able to interpret their production data using data mining techniques and share that data through the AQUASMART open cloud.

This paper is structured as follows: Section 2 presents the related semantic approaches. The overview and architecture of the project that serves as validation for the ontology are explained in section 3. Section 4 presents the AQUASMART semantic referential. Section 5 presents the IT semantic services developed, the tools used and the interdependencies between components that implement the AQUASMART functionalities. Finally, section 6 presents our main conclusions and additional consideration on future work.

2 RELATED WORK

Big Data analytics and knowledge extraction from the aquaculture domain, is something relatively new. This chapter address some relevant works that have been done relatively to the technology and the domain.

2.1 ICT in Aquaculture

Recent ICT projects on the aquaculture domain, aim to ease the access to information about scientific farming practices and market prices through web portals (De et al., 2008), while others propose innovative and unique models of information exchange on farm history, crop details, soil details, weather data, farmer details, case sheets, photo bank and a library (Vimala, 2009). The multilingual issue is addressed by (Mondal et al., 2011), where an online tool provides answers to questions asked by farmers and agro-professionals over the Internet.

Also, a very important step to adopt ICTs in Aquaculture environments is to reach the end-users in an

early-stage, even during the learning process. In (Seixas et al., 2015) the authors review educational means used in teaching and learning in the area of aquaculture, fisheries and aquatic resources management at European level, with specific consideration on the use of ICT and e-learning tools. It concludes that there is a real and urgent need to "train the trainers" to use ICT in their teaching environments. Additionally, from the students' end, there is a strong desire to learn more about the application of e-learning tools and to use them in their learning process.

2.2 Ontology Development

The method proposed in our work, adopts an association rules learning technique in order to discover relevant relations among key terms in a document corpus, and additional human input to perform the mappings between terms (frequent item sets) and ontological concepts and the establishment of the final scores on each relation. In simple words, frequent item sets are groups of items that often appear together in the data.

This will reuse the findings proposed by (Paiva et al., 2013) and extend it to the aquaculture domain. Figure 1 depicts the methodology to be addressed by this work regarding thesaurus building.

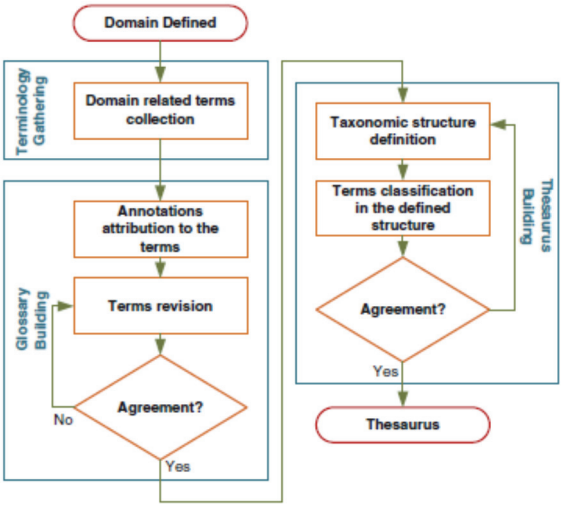

Figure 1: Thesaurus building methodology (Sarraipa et al., 2010).

One of the focus of the work addressed where, is the development of a semantic referential for the aquaculture domain. Regarding ontology development, the scientific literature is vast and addresses several approaches and tools for dealing with its specification and management. Several efforts have focused on extracting structured ontologies from unstructured text.

2.3 Multilingual Data

The introduction of an innovative multilingual knowledge base capacity suitable for the Aquaculture sector, which would enable large volumes of data to be accessible as semantic interoperable data and knowledge will improve significantly the sector and ultimately the EU's competitiveness.

Three tools to support the multi-linguist requirements: 1) a tool for language independent document representations (Canonical Correlation Analysis (CCA) mappings) (Kuss and Graepel, 2003); 2) a tool for multilingual semantic annotation (Enrycher) (Jožef Stefan Institute, 2012); and 3) a statistical machine translation tool (MOSES) (Koehn et al., 2007). They will be used to support the ontology mapping processes which involve ontology concepts and annotated documents in several languages.

Enrycher is a service-oriented system providing shallow as well as deep multilingual text processing functionality at the document level. It can be used to perform multilingual semantic annotation on documents in common open knowledge bases such as DBpedia (Lehmann et al., 2015), and can support the local to reference ontology mapping processes.

The MOSES is an open source statistical machine translation system to translate documents stored in local databases across several language pairs in order to share the knowledge between Aquaculture companies. The system will also be used to support the multilingual training and learning platform by translating training materials such as best practices documents and video lectures. Linguistic resources such as the Fisheries Glossary can be incorporated in the process (Countryside Council for Wales, 2001) by using it as a parallel corpus. Having large collections of short documents can be used to adapt the language model for the target language.

3 AQUASMART CONCEPT

The data collected in the AQUASMART open data cloud is suitable to be reused in other industrial domains if needed, (e.g., environmental or transportation data), providing a cross-sectorial setting to the provided solution. The cloud is enriched with a layer of multilingual information (multilingual mappings and multilingual linguistic information) and with a set of services for creating, representing and accessing that multilingual information.

The prime goal is to accelerate innovation in Europe's Aquaculture through technology transfer for

the deployment of an open data solution through multilingual data collection and analytics solutions and services, turning the large volumes of heterogeneous aquaculture data that is distributed across the value chain, into an open cloud of semantically interoperable data assets and knowledge. Each such systems usually uses different data representations, using its native languages and knowledge organization tools such as vocabularies and classification systems to manage and organize information. Although the practice of using knowledge organization tools to support document tagging (e.g. thesaurus-based indexing) and information retrieval (e.g. 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 project technology transfer goals are to take the state of the art in multilingual data collection tools, analytics solutions and services, semantic interoperability methods and data mining procedures to implement a global open aquaculture data information system, able to respond to specific user needs and profiles Figure 2.

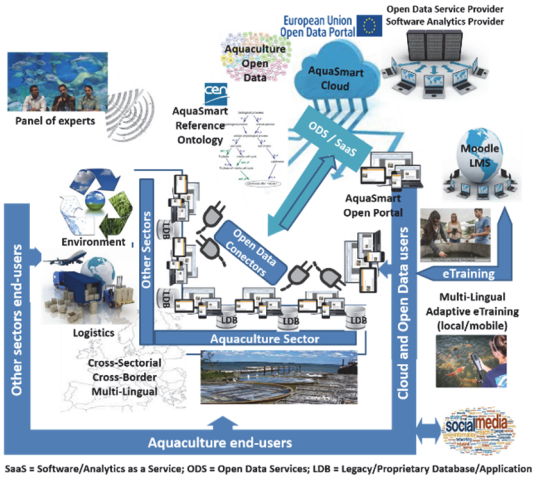

Figure 2: AQUASMART Concept.

Actually, aquaculture companies have never tried to transform data that are captured into knowledge, and share this knowledge to improve efficiency, increase profitability and do business in a sustainable, environmentally friendly way. In other words, data are just captured but not exploited. AQUASMART represents a big innovation in this direction, it adds the dimension of global open data. The improvement and the innovation become even bigger because the quality of the analysis that will be performed by the companies will be dramatically improved. Hence, even the exchange of information itself is a huge business innovation. It is the first time that companies will be able compare their results to the ones of other companies and benchmark their performance. This is going to create a multiplier effect and boost competition for improvement in the sector. This will be further facilitated and enhanced by the integration with the European Union Open Data Portal and the exchange of information in the social media.

AQUASMART vision relates to implementing a state of art multilingual open data framework that companies can use to seamlessly access global data and take more knowledgeable decisions using multilingual information. H2020's vision suggests that enterprises must move away from silo solutions, used behind the closed doors of company, to a more open data technological solutions built for the industrial sector to enhance their operations. However, the actual state of practice is that knowledge is transferrable and sharable but with significant barriers for semantic compatibility.

The main mission is driven by the business need of the European aquaculture companies, when companies have business objectives that they cannot achieve due to lack of instruments that would enable them to manage and access to global knowledge and big data, in a multi-lingual, multi-sector and cross-border setting.

4 SEMANTIC REFERNCIAL

To improve the ability of aquaculture companies to innovate across their value chain, there is a need to provide multi-lingual data, which must be interoperable through various products or services. To achieve this objective, the proposed approach integrates a reference ontology that supports the integration of heterogeneous data (including multilingual) that concerns the aquaculture environment.

The proposed semantic referential provides the distributed composition, formality, richness and quality of information required among the aquaculture sector to ensure that all the actors within the production process "speak the same language".

The creation of the semantic referential followed a method for designing and developing a domain ontology with inputs from knowledge experts, providing the necessary insights towards the improvement of the efficiency of the aquaculture production processes. Such experts, contributed with their knowledge about the aquaculture production, the actors involved, and the data generated during the production process (Oliveira et al., 2015).

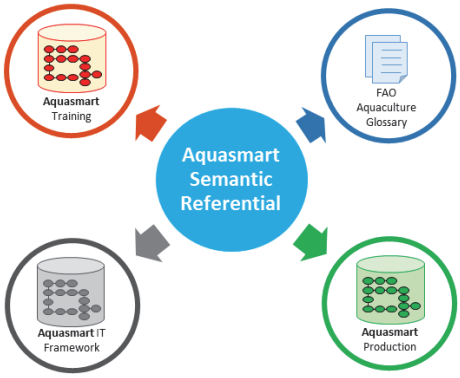

Figure 3: AQUASMART Semantic Referential.

This semantic referential, as seen in Figure 3, is composed by four main areas, (i) the AQUASMART production domain; (ii) FAO aquaculture glossary; (iii) AQUASMART training modules and; (iv) the AQUASMART IT framework. In the following sections, the authors address each of these modules, which comprises the overall AQUASMART semantic referential. Each of the modules are being integrated under the same umbrella (AQUASMART semantic referential) and formalized in OWL language.

4.1 AQUASMART Production Domain

In the AQUASMART context, knowledge experts are the end users (mainly fish farmers). The purpose of involving such experts in the process, is not only to provide input to the semantic referential, but also to perform a quality review of the AQUASMART training courses. With the help of these experts, the main structure of the AQUASMART ontology was developed to accommodate all the important and necessary information that will support all the project services and functionalities.

The ontology is mainly separated in two concepts, the "Aquaculture Production Entities" and the "Grow Out Data Analysis". The first contains the all the aquaculture related entities, while the second one contains the key performance, the process and production related data.

The "Aquaculture Production Entities", are the main components of the production operation, including actors involved in the process, species being produced (e.g. seabream and seabass), location of the fish farm, and the type of cages used to store the fish. The "Grow Out Data Analysis", is focusing on process steps, production data and indictors. A sample of relevant production parameters is shown on Table 1.

Table 1: Production Parameters.

Geographical Region	Species
Hatchery	Broodstock origin
Hatchery Quality (text)	Stocking Month
Average Weight	Mortality
Avg. Temperature	Avg. SFR
Avg. Fish Density	Oxygen

This ontology will be integrated with an aquaculture glossary that contains the aquaculture related terms and their definitions. To make available fully interoperable multi-lingual data products and services in the Aquaculture, AQUASMART makes use of the FAO glossary for aquaculture (FAO, 2015).

4.2 FAO Aquaculture Glossary

From the AQUASMART perspective, the multilingual data generated within the aquaculture domain, can be exploited as a layer of services and resources by seamlessly adding (i) linguistic information for data and vocabularies in different languages, (ii) mappings between data with labels in different languages, and (iii) services to dynamically access and traverse linked data across different languages.

We envisage a multilingual aquaculture where an end-user would query the "Aquaculture Open Data Cloud" in his/her own language, and would get the relevant data in that language.

The primary objectives of the AQUASMART glossary is: (i) to serve as a reference to fish farmers, consultants, administrators, policy makers, developers, engineers, agriculturists, economists, environmentalists and any other actor interested in aquaculture; (ii) and to facilitate communication among experts and scientists involved in aquaculture research and development.

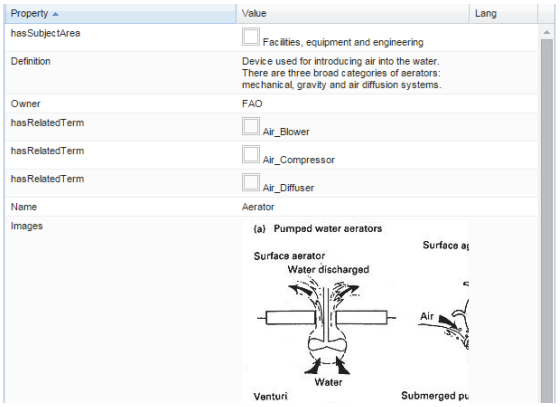

Figure 4: Glossary Term Example.

The glossary supports a multi-lingual approach that includes, in an initial phase, English, French,

Greek, Hebraic, Spanish and Portuguese terms. Each term has properties which define it (Name, definition, related term, synonyms, subject area, translation, and image). This kind of information is what defines and gives semantic meaning to terms. Figure 4 shows an example of a glossary term.

4.3 Ontology Training

The training development methodology translates design specifications into training materials. The methodology presented by Sarraipa et al., (2013) starts by identifying the training objectives and the target audience including desired roles & competences. Then it uses an appropriate instructional approach to perform the training courses' materials development, complemented with a set of different quality reviews.

The overall process of developing training follows a specific process, composed by three different task tracks (training development, overall training validation and training execution) that complement each other. The training development track starts by defining the course's synopsis according to the directives obtained from a training overall objective. Thus, it was identified the need of defining all these objectives in a courses synopsis. A course synopsis is an official description of the course as stated in the institution's catalogue of courses. It should indicate the overall goal of the course, briefly characterize the main topics covered, point out why the course is important to students, identifying any special instructional methods to be used, and comment on what background students should have in order to best appreciate the course content. The courses synopses also act as guidelines to the training course's authors.

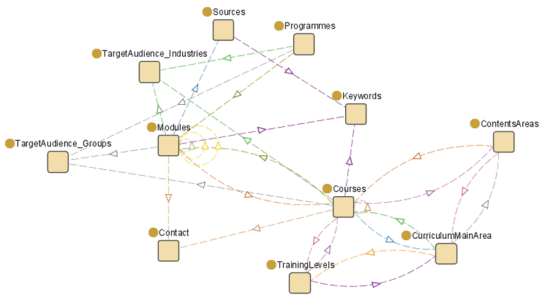

Figure 5: Training Ontology.

The AQUASMART ontology will also be used, to represent the training KB (Knowledge Base), facilitating the categorization of its elements and subsequently reasoning over it. It should contribute to the skills and competencies development of the trainees as required for specific understanding and exploitation. This reflects the need to develop, organise and run courses, for example, to train "future users", in how to use a specific software or extracting relevant patterns from aquaculture production data. Figure 5 presents the relations between training concepts built in Protégé (Stanford Center for Biomedical Informatics Research, 2012).

In this model each learning Module has several concepts associated, the Sources concept contains information about the sources referred in the Module, Contact includes the contact information of the author of a Module or Course and Keywords that contain a list of all relevant keywords needed for describing the contents of the Module. A Course, other than Contacts and Modules that contain the course also includes Keywords (that include Keywords inherited from its Modules) and belongs to a Curriculum Main Area that is divided by Content Areas and Learning Levels. Each Module and Course has a Target Audience Group and a Target Audience Industry, to be recommended accordingly to the profile of the learner. Finally, a pre-defined Programme is defined for a specific Target Audience Industry and Target Audience Group.

5 AQUASMART SERVICES

The AQUASMART IT framework is not yet developed at the current stage of the project. It will be in this framework that the AQUASMART services will persist.

One of the services is the ability to search for Aquaculture partners through the ontology. The ontology will support a search feature that allow users to find suitable companies in Aquaculture domain to partner. This feature allow to search companies by different criteria like water temperature, type of fish produced, size of production, country, this are some examples supported.

Providing the achieved results to other users is a service that is being worked on. The main idea of this service is to provide the results obtained during the project. This can be in text, formulas or other type of format that can be understand by users external to the project.

5.1 Knowledge Search Engine

The other service developed is the training search engine to support the Aquaculture domain. Training materials are a necessary part of any program or activity that involves knowledge acquisition and retention (Wikihow, 2015). For this, authors found appropriated to define "oLEARCH - Ontologies LEArn by

seaRCHing", as a new concept related to ontologies able to change/adapt their knowledge (to learn) through their users' patterns of searching/reasoning. The concept was inspired from the concept LEARCH defined by (Ratliff et al., 2009) that means "LEArning to seaRCH" and was defined to represent algorithms for imitation learning in robotics with the main purpose to search something.

oLEARCH is a training materials search engine application available to users by Internet. This system learns from user's searched training materials concepts improving the KB.

The oLEARCH function uses an algorithm supported in an instance-based learning approach based on user interactions. In instance-based learning, training examples are stored verbatim, and a distance function is used to determine which member of the training set is closest to an unknown test instance (Witten and Frank, 2005). In oLEARCH, such distance function is represented by the semantic distance, which is the inverse of the semantic relatedness between the users introduced concepts and the training materials classified in the reference ontology. Thus, oLEARCH provides to the users a set of training materials that are close to their introduced concepts in terms of semantic relatedness. Then, users are able to select the most appropriated training materials from this set of possible choices. These last users' selections are also used, as a last feedback, to increase the semantic relatedness weight of the selected training materials associated concepts.

6 CONCLUSIONS

The work described within this paper relates to the development of a domain ontology, able to support and describe the analysis of aquaculture production data, but also, the training and IT services which composes the AQUASMART platform. Although final conclusions are not yet validated, preliminary analysis led us to conclude that the Aquaculture domain is lacking for semantic approaches which enable data understanding intra and inter-organizations. The formalization and validation of a common semantic reference model, which is able to drive new and dynamic collaborations between aquaculture companies and consequently generate new business opportunities to them, can be seen as first step towards semantic interoperability. From an application scenario perspective, the objective of the AQUASMART semantic referential will enable the understanding of data analytics resulting from the production data. The idea is to semantically annotate the results of the correlations found in batches of production data, with ontology concepts in order to give meaning to data analytics results. Multilingual is another important feature due to the fact that knowledge transfer is one of the main challenges to be addressed here.

With the proposed approach presented here, there will be opportunity for innovation in the aquaculture industry such as transforming data into global knowledge, and use this knowledge to improve efficiency, increase profitability and do business in a sustainable, environmentally friendly way; Better and perfect view of the life to date fish behaviour and the living inventory (biomass) that exist in a farm, based on the analysing of all environmental and biological data that will exist in the local system and at global level. By knowing the global parameters that affect the production, the companies will be able to make accurate estimations of the growth of the fish and the result of the production every day.

ACKNOWLEDGEMENTS

The authors acknowledge the European Commission for its support and partial funding and the partners of the research project: H2020-644715 AQUASMART.

REFERENCES

Countryside Council for Wales, 2001. *A glossary of Marine Nature Conservation and Fisheries.* [Online] Available at: http://jncc.defra.gov.uk/pdf/glossary.pdf

De, H. K., Saha, G. S., Srichandan, R. & Vipinkumar, V. P., 2008. New initiatives in fisheries extension. *Aquaculture Asia,* Volume 13, pp. 16-19.

FAO Fisheries and Aquaculture Department, 2014. *FAO Global Aquaculture Production Volume and Value Statistics Database Updated to 2012,* s.l.: s.n.

FAO, 2015. *FAO Glossary.* [Online] Available at: www.fao.org/faoterm/collection/aquaculture/en/[Accessed 2015].

Jožef Stefan Institute, 2012. *Enrycher.* [Online] Available at: http://enrycher.ijs.si/

Koehn, P. et al., 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* Prague, Annual Meeting of the Association for Computational Linguistics (ACL).

Kuss, M. & Graepel, T., 2003. *The Geometry Of Kernel Canonical Correlation Analysis,* s.l.: Max Planck Institute for Biological Cybernetics.

Lehmann, J. et al., 2015. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal,* 6(2), p. 167–195.

Lima, C., Zarli, A. & Storer, G., 2007. Controlled Vocabularies in the European Construction Sector: Evolution,

Current Developments, and Future Trends. In: *Complex Systems Concurrent Engineering.* s.l.:Springer London, pp. 565-574.

Mondal, P. P., De, H. K., Saha, G. S. & Radheyshyam, 2011. Information and Communication Technology (ICT) and Aquaculture. *Aquaculture International,* pp. 32-35.

Oliveira, P. et al., 2015. *A Knowledge-Based Approach for Supporting Aquaculture Data Analysis Proficiency.* Houston, Texas, USA, s.n.

Paiva, L., Costa, R., Figueiras, P. & Lima, C., 2013. *Discovering Semantic Relations from Unstructured Data for Ontology Enrichment - Association rules based approach.* Lisbon, IEEE.

Ratliff, N., Silver, D. & Bagnell, J. A., 2009. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots,* 27(1), pp. 25-53.

S. R. Ray, A. T. Jones, 2006. Manufacturing interoperability. *Journal of Intelligent Manufacturing,* Volume 17, pp. 681-688.

Sarraipa, J. et al., 2013. *E-TRAINING Development Approach for Enterprise Knowledge Evolution.* San Diego, CA, USA, s.n.

Seixas, S., Dove, C., Ueberschar, B. & Bostock, J., 2015. Evaluation on the use of e-learning tools to support teaching and learning in aquaculture and aquatic sciences education. *Aquaculture International,* pp. 825-841.

Stanford Center for Biomedical Informatics Research, 2012. *Stanford's Protégé Home Page.* [Online] Available at: http://protege.stanford.edu/[Accessed 3 September 2012].

Vimala, D., R., T. M. P., A. K., 2009. An innovative information and Communication technology for transfer of technology for aquaculture. *Aquaculture Asia,* Volume 14, pp. 23-24.

Wikihow, 2015. *http://www.wikihow.com/Develop-Training-Materials.* [Online] Available at: http://www.wikihow.com/Develop-Training-Materials[Accessed 2015].

Witten, I. H. & Frank, E., 2005. *Data mining: practical machine learning tools and techniques.* 2nd ed. S.l.:Morgan Kaufmann Series in Data Management Systems.