Web mining to create semantic content: A case study for the environment

Georgia Theocharopoulou and Konstantinos Giannakis

Ionian University, Department of Informatics,
Tsirigoti Sq. 7, 49100 Corfu, Greece
{c11theo,c11gian}@ionio.gr

Abstract. In this study, the goal is multifold. At first we present a summarized review of terms and facts regarding the branch of ecoinformatics, web mining and the semantic web. In Section 2 we provide some related work derived from the current literature upon the web mining and the production of semantic content. The main part of our work follows presenting a notional model for building semantic content through 2-level web mining. This is achieved in web sites containing environmental data. We conclude mentioning the importance of this contribution from different points of view.

Keywords: web mining, semantic web, ecoinformatics

1 Introduction

Many research projects have been focused on the application of data mining techniques to the WWW[1], which is referred to as Web mining. The term Web Mining can be broadly defined as the discovery and analysis of useful information from the Web [5]. With the explosive growth of information resources available online, there is a certain difficulty in finding relevant information. When trying to find specific information using a keyword query, usually it depends on the keyword that was used to find responses with high precision. The use of Semantic web helps to create new knowledge out of the information available on the Web. According to the tools of mining that are used there are referred three types of Web mining.

Web Content mining. Mining, extraction and integration of information from the Web contents such as textual, image, audio, video, metadata as well as hyperlinks. Most of these data is unstructured text data. There are also semi-structured data and a more structured data such as data in the tables or database generated HTML pages. There are various techniques which aim to process unstructured (textual) information, in order to extract meaningful content, such as knowledge discovery in texts (KDT). Web content mining is referred to be

[1] World Wide Web

dfferentiated in two dfferent points of view: IR (Information Retrieval) and DB (Database) views. The goal of IR is retrieving information, categorization and filtering the information. However the retrieved objects might be inaccurate and the main reason for this, is that information retrieval deals with natural language text which could be semantically ambiguous. A database retrieval (DB) system tries to model and integrate data that has a semi defined structure.

Web Structure mining. There are techniques that are used to discover the patterns which can model the link structures of the Web. These models can be used to extract useful information, such as web page categorization, interesting Web structures and quality reports from different Web sites.

Web Usage mining. Extracts useful information from the data collected from Web servers. The aim is to generate meaningful patterns and relationships from primarily semi-structured data, stored in Web and applications server access logs. The goal is to model the behavioral patterns and generate the profiles of users. Knowledge about a user's behavior can lead to future applications where the structure and content is based on the user's pattern profile.

Let's consider why somebody would want to process online data. First of all, being able to structure the amount of information available in the World Wide Web accommodates understanding and sharing of common data. Knowledge discovery, not only extracts useful knowledge from databases, but also enables the reuse of this knowledge. From 1999, when Tim Berners-Lee expressed the vision of the Semantic Web there has been a lot of development in web mining in order to use intelligence as an extension. The evolution of the existing Web to a Semantic Web uses semantic applications, such as RDF[2] and ontology description languages. Ontologies are key tools to Semantic Web because they offer an opportunity to signicantly improve knowledge management capabilities. Ontological analysis clarifies the structure of knowledge and separates domain knowledge from the operational knowledge.

One of the most common goals in developing ontologies is enabling common understanding of the structure of information among people or software agents. The basic problem is the same: making sense of already existing data in order to be classified in conceptual structures. Lack of shared understanding leads to poor communication and limited or incomplete sharing of information. For example, unlike data collected from molecular biologists, data collected by environmental biologists, are rarely made available to many other scientists [15]. In recent years the use of ontologies in the domain of biology and bioinformatics has proliferated. In molecular biology there exist many databases for storing data and some using identical labels and categories but with a different meaning. Ontologies can bridge the different notions in various databases and provide systematically a semantic repository in this application domain.

[2] Resource Description Framework

Ecological science studies hypotheses based on conceptualization of living organisms and their relationship with the environment. Variables of interest to ecologists often include concepts with ambiguous interpretation. Developing ontologies to control and clarify terms in ecology, can enhance data sharing and analysis [12]. Ecoinformatics involve projects that apply knowledge representation and Semantic Web technologies to problems in discovering and integrating ecological data and data analysis techniques. These technologies rely on ontologies that appropriately capture and encode scientific knowledge from the domains of interest [37].

The same situation with multiple operational denitions of terms and variable meanings in different contexts, extends to other interdisciplinary fields as well, such as Environmental Science, that integrates physical and biological sciences. Environmental Science is not limited to the study of the environment, on subjects like understanding physical, chemical, and biological processes, but also on interactions and effects of environmental problems, like the global climate change.

Human impacts upon natural systems requires access to multi-disciplinary information, including chemical, behavioral, geological, meteorological and sociological data [12]. For instance, the individual and household energy consumption and carbon dioxide energy emission is estimated to be forty percent of the total amount for the U.S. Policymakers lack of access to the key findings of behavioral and social science studies on household energy behavior [22]. Studies from across western industrialized societies have shown that the majority of people have favorable environmental attitudes but they are not always translated into appropriate behavior [23]. Decision making is limited by incorrect or incomplete information. Individuals often act in ways that they believe to benefit the common good, when in fact these actions are counterproductive due to inability to process information [22].

Large amounts of environmental data are increasingly available on the World Wide Web. Finding data relevant to a certain query might be a difcult task, since most search engines use text-matching algorithms, which return large numbers of results that have to be manually examined in order to decide the relevance to the field of query. Ontologies can efficiently organize all these information by effective management representation of knowledge.

2 Related work

Web mining techniques can be used to solve the information overload problem. Web Mining could be decomposed into the following subtasks: 1. Resource finding 2. Information selection and pre-processing 3. Discovering general patterns 4. Analysis of the mined patterns [17]. Many Web based applications developed the past few years, are using these web mining techniques. CiteSeer, for example, is one of the most popular online bibiliographic indices related to Computer Science. CiteSeer works by crawling the Web and downloading research related papers. The key element of CiteSeer is the Automatic Citation Indexing.

Currently the web is mainly a collection of unstructured or semi-structured data. Semantic Web enables intelligent access to distributed information, providing efficient information retrieval and knowledge discovery to meet the user needs.

The most public usage of Semantic Web technologies is the website for the British Broadcasting Corporation (i.e., the BBC). In 2010, their entire World Cup website was powered by Semantic Web technologies, as was reported on ReadWriteWeb and SemanticWeb.com. Today, large portions of their public website are run on Semantic Web technologies. Creating web identifiers for every item the BBC has an interest in, enables very rich cross-domain user journeys. This means BBC content can be discovered by users in many different ways [27]. Time Inc., Elsevier, and the Library of Congress all also have production systems built using Semantic Web technologies.

A Semantic Web vocabulary can be considered as a special form of an ontology. Interest in developing ontologies for describing ecological data has emerged due to the fact that there exists a flood of data in ecological and environmental sciences which lack of formalization. Recently, have been developed projects such as "Science Environment for Ecological Knowledge" (SEEK) and "Semantic Prototypes in Research Ecoinformatics" (SPIRE), which apply data analysis techniques with knowledge representation and Semantic Web technologies to discover and integrate ecological data. These technologies rely on ontologies that encode knowledge from domains of interest. The Extensible Observation Ontology (developed under the SEEK project), describes basic concepts and relationships for observational datasets, including eld, experimental, simulation and monitoring data. The ETHAN Evolutionary Tree ontology is an OWL-based representation of the Animal Diversity Web data [28], which also can be used to describe the evolutionary relationships among organisms from the Integrated Taxonomic Information System ITIS [29].

DBpedia is a project which aims to represent knowledge by extracting structured content from the information created as part of the Wikipedia project, which is then made available on the World Wide Web. Users are allowed to query relationships and properties associated with Wikipedia resources, including links to other related datasets. This knowledge base supports complex semantic queries.

3 Mining the web

So far we have referred some definitions and terms regarding semantic web and environmental issues and we presented mining methods that are currently implemented.The main part of our study follows in this section. Obviously, the mining procedure will occur on the content of the web sites, i.e. it's a text web mining function in sites containing rich datasets. These sites can be wikis, databases etc. having as a leading attribute the wealth in data, the interlinking among them and the fact that they offer many, expertised details about their contents.

During our investigation across the web for data sources satisfying the above requirements, we concluded that there is quite a huge number of candidate sites.

Many of these sites owe their data wealth to several projects run by institutions, research groups or universities worldwide. In addition to their richness and attention to details, they have one enormous advantage: they are built by experts of each field. An additional, as well as, essential property of these data sources is the fact that almost all of them consist of structured or semi-structured data assisted by almost perfect interlinking, so as the mining process is accommodated and the need for a supervisor over the mined results declines. In sites such as [34] and [36] the information is structured in an sensational way in order that a machine can manipulate their content.

Our attention is focused on the Encyclopedia of Earth. It's the first important step in our proposed conceptual model. Using data derived from there, we could build the first level of categorization. Nevertheless, we propose an hybrid model that combines the general Encyclopedia of Earth with more detailed and expert on-line databases (2nd level of categorization) in order to achieve better performance and results. As far as the storage is concerned along with the processing of the derived data, its evident that storage and manipulation of *big data* nowadays is easier than ever before. There are plenty of tools, both in hardware and software level offering vital assistance towards this direction. Furthermore, the nature of the data in our concept is such so that their process can be done in a parallel mode.

Although the terms "classification" and "clustering" associated with machine learning algorithms are different (the former has to do with already known classes-categories whereas in the latter the classes-categories are self-formed during each training stage of the algorithms), for the rest of the paper, it will be meant the same thing to avoid misunderstanding. Since the categories will be mined as well, the term "clustering" might be more appropriate.

Next, we proceed in showing more technical details explaining our way of thinking. In next figures, we make clear our concept, i.e. the integration of different rich web data bases with ultimate goal the production of semantic material. In Fig. 1 we present the abstract model of DBpedia and the analogy of our proposal. DBpedia is based in Wikipedia articles (and specific segments of its articles, such as infoboxes). In comparison to that, we introduce a union of different parts to form the whole frame to mine. It is clear now that we use multiple sites to accomplish the intended web mining, having as a centralized hierarchical source the [31], with the rest of the specialized sources enriching the whole ontology and its vocabulary with vital details.

Next, in Fig. 2 we explain how the incorporation described above takes place. The "hub" of our abstract model is the Encyclopedia of Earth, which acts like the first level classifier. Afterwards, the different entities derived from the Encyclopedia are enriched and increased by other data source in the web, such as the FishBase about fishes.

In the last figure (Fig. 3) we show the dendrogram that represents the hierarchy in the clustering process. It's indicative and it shows again the disjoint

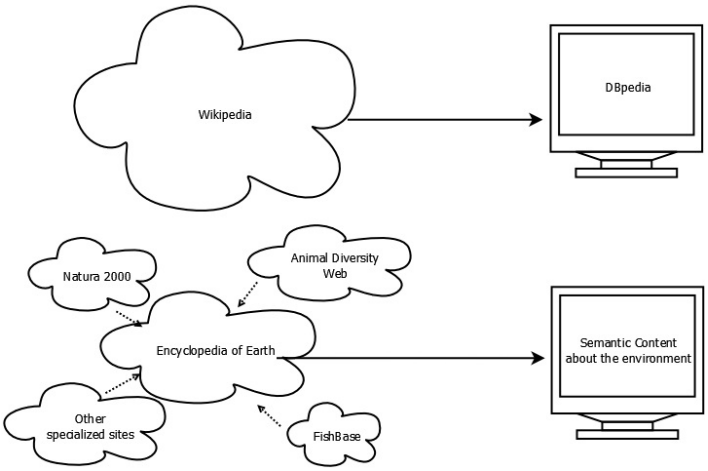

Fig. 1. The analog of our concept to DBpedia/Wikipedia.

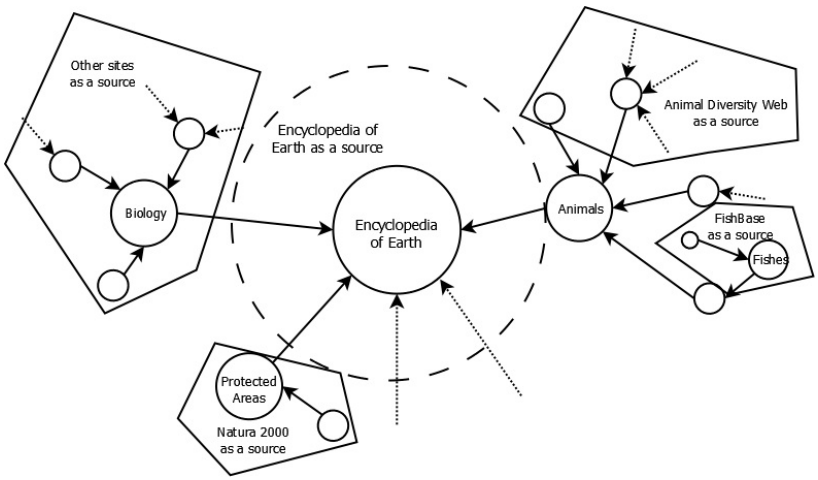

Fig. 2. Encyclopedia of Earth as a hub integrated with other data sources.

into two levels, as previously mentioned. We should point out that a serious disadvantage of our approach is the lack of a common data source, resulting in the necessity of different ways of handling each dataset of every sites.

This treelike structure can be seen as a "decision tree". Any data derived from any source will be analyzed and then classified in one of the entities/branches of the tree. The first level is the derivation of the upper categories, a course implemented by mining the Encyclopedia of Earth. The leaves and the lower level nodes of the tree consist of the derivation of multiple data of every data source according to its subject e.g. FishBase will undertake the duty of classifying the harvested instances of animals that belong to the category "Fishes".

The classification procedure can be implemented in various ways. There are plenty tools to grab the data in the first place and several algorithms to complete the second stage of the whole procedure. One such efficient algorithm is ROCK [10] that clusters the data in groups taking advantage of the interlinks among them (crawlers can deploy and explore link by link).

4 Necessity of this framework

In this section we explain why this framework is important and how someone can use the results of the procedure. First of all, enhancing the WWW with semantic data related to earth can be found usefull. Having published content with semantic contain empowers the query spectrum. More sophisticated and targeted queries, without the need of complex and, often, redundant information can be posed and answered. Also applications with vital meaning for mankind, such as fire protection, pollution reduction etc. can offer better results taking advantage of the smart web.

Other scientific and humanistic fields, like tourism, agriculture, weather forecasting and so on, can utilize smart information about the environment. For example, farmers can find useful information about plants and pesticides by posing simple but meaningful queries upon semantic environmental data. Furthermore, the linking of our results to other ontologies and/or Linked Data published on the web (e.g. DBpedia) can be established.

5 Future work and conclusion

Since our proposed model lies in an abstract level, as a future work we intend to put it into practice. Having in mind the continuous dynamic import of new data on the web, we plan to keep searching for new supplementary data sources to enhance the performance of this framework. This can be done using intelligent agents based on our study's results, a case to be examined in a future work. Finally, we should report that an application of our conceptual framework can and is encouraged to be linked with already existed ontologies. For example Open Linked Data is a wealthy area and can adopt a full and robust frame regarding environmental essences.

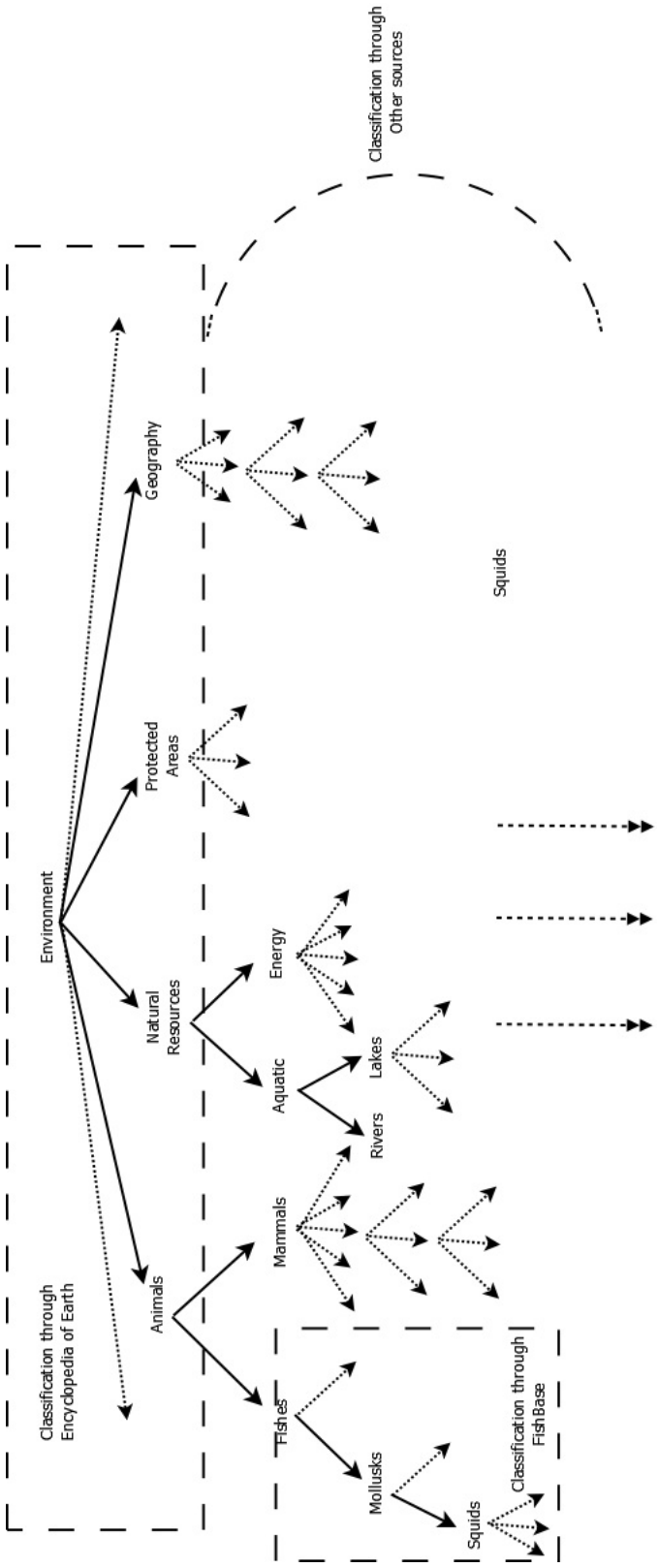

Fig. 3. A tree-like representation of the hierarchies along with their sources.

In this work we presented a synoptic review of the association between environment and computer science. We conclude this study noting the commonly accepted point of view that environmental issues need to be carefully examined from every point of view, including data mining and computer science. In this direction, semantic web, the future trend of internet, can and should be used in the context of environmental facts. The WWW is rich of earth information which our conceptual model tries to harvest and exploit in an efficient, fast and automated way.

References

1. Koursaris, S., Giannakis, K.: Hellenic Natural Resources with Ontologies. In Local proceedings of 15th Panhellenic Conference of Informatics, 2011, Kastoria, Greece (2011)
2. Auer, A., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, I.: DBpedia: A Nucleus for a Web of Open Data, In The 6th International Semantic Web Conference (ISWC 2007) Busan, Korea, (2007)
3. Kobilarov, G., Bizer, C., Auer, A., Lehmann, J.: DBpedia-A Linked Data Hub and Data Source for Web and Enterprise Applications. In WWW2009, Madrid, Spain (2009)
4. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., M. Smethurst, M., Bizer, C., Lee, R.: Media meets semantic web-how the bbc uses dbpedia and linked data to make connections. In 6th European Semantic Web Conference (ESWC2009), Semantic Web In-Use Track, (2009)
5. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. In roceedings of the 9th IEEE International Conference on Tools with Articificial Intelligence (ICTAI 97), (1997)
6. Russom, C.: Mining environmental toxicology information: web resources. Toxicology 173, 75-88, Elsevier Science Ireland Ltd (2002)
7. Zweigle, O., H$\ddot{a}$ussermann, K., K$\ddot{a}$ppeler, U. P., Levi P.: Extended TA Algorithm for Adapting a Situation Ontology. In: J.-H. Kim et al. (Eds.): FIRA 2009, CCIS 44, pp. 364371. Springer, Heidelberg (2009)
8. Getoor, L., Diehl, C. P.: Link Mining: A Survey. In ACM SIGKDD Explorations Newsletter, Vol. 7, issue 2, pp 3-12 (2005)
9. Melli, G., Ester, M.: Supervised Identification and Linking of Concept Mentions to a Domain-Specific Ontology. In CIKM10, Toronto, Ontario, Canada (2010)
10. Guha, S., Rastogi, R., Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes. In Proceedings of the 15th Int. Conf. on Data Engineering (2000)
11. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154165, (2009)
12. Madin, J.S., Bowers, S., Schildhauer, MP., Jones, MB.: Advancing ecological research with ontologies. Trends in Ecology and Evolution 23(3), 159–68 (2008)
13. Villa, F., Athanasiadis, I. N., Rizzoli, A. E.: Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. Environmental Modelling Software, vol. 24, issue 5, 577-587 (2008)
14. Freitas, F., Stuckenschmidt, H., Noy, N. F.: Ontology Issues and Applications. Journal of the Brazilian Computer Society, vol.11, no2 (2005)

15. Williams, R. J., Martinez, N. D., Golbeck, J.: Ontologies for ecoinformatics. Web Semantics: Science, Services and Agents on the World Wide Web, vol.4 Issue4, (2006)
16. Sachs J., SimsParr C., Parafiynyk A., Lushan R. P., Ding H. L., Finin T.: Using the Semantic Web to Support Ecoinformatics. In: AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition (2006)
17. Kosala R., Blockeel H.: Web Mining Research: A Survey. In: ACM SIGKDD Explorations Newsletter, Vol.2, Issue1,(2000)
18. Watson M.: Practical Semantic Web and Linked Data Applications. E-Book (2011)
19. Lepczyk C.A. , Lortie C. J. , Anderson L. J.: An ontology for landscapes. Ecological Complexity, vol.5,Issue3, (2008)
20. Bharanipriya, V., Kamakshi, V. P. Web content mining tools: a comparative study. International Journal of Information Technology and Knowledge Management, Vol. 4, No. 1, pp. 211-215, (2011)
21. Jannike, P., Jsendal V.: The Semantic Web from a humanities perspectives. Master Thesis
22. Carrico, A. R., Vandenbergh, M. P., Stern, P. C., Gardner, G. T., Dietz, T., and Gilligan J. M.: Energy and Climate Change: Key Lessons for Implementing the Behavioral Wedge. Journal of Energy and Environmental Law, Vol. 1, (2010)
23. Brandon. G., Alan lewis. Reducing household energy consumption: a qualitative and quantitative field study. Journal of Environmental Psychology, 19, 7585, (1999)
24. Jannike, P., Jsendal V.: The Semantic Web from a humanities perspectives. Master Thesis
25. Srivastava, J., Desikan, P., Kumar, V.: Web Mining - Concepts, Applications Research Directions. Studies in Fuzziness and Soft Computing, Vol. 180, pp. 275307, (2005)
26. Chakrabati, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Rghavan,P., Rajagopalan, S. and Tomkins, A.: Mining the link structure of the World Wide Web. IEEE Computer, 32(8), pp.60–67, (1999)
27. Case Study: Use of Semantic Web Technologies on the BBC Web Sites, http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/
28. Myers, P., R. Espinosa, C. S. Parr, T. Jones, G. S. Hammond, and T. A. Dewey. 2006. The Animal Diversity Web (online). Accessed April 29, 2012, http://animaldiversity.ummz.umich.edu
29. Retrieved April 29, 2012, from the Integrated Taxonomic Information System (ITIS), http://www.itis.gov
30. Generic Model Organism Database project, http://www.gmod.org/wiki/Main_Page
31. The Encyclopedia of Earth, http://www.eoearth.org/
32. Semantic Web for Earth and Environmental Terminology (SWEET), http://sweet.jpl.nasa.gov/
33. Applying semantic web technologies in research ecoinformatics, http://spire.umbc.edu/us/
34. Environmental Wiki, http://www.envirowiki.info/Main_Page
35. DBpedia, http://dbpedia.org/About
36. Froese, R. and D. Pauly. Editors. 2011.FishBase. World Wide Web electronic publication. www.fishbase.org
37. Wikipedia. http://en.wikipedia.org/wiki/Main_Page