A taxonomic service for species identification

Daniel Fuentes, Nicola Fiore
Doñana Biological Station (Spanish National Research Council), Seville, Spain
dfuentes@ebd.csic.es
Ecology Unit, Università del Salento, Italy
nicola.fiore@unisalento.it

ABSTRACT

Taxonomy is the science of discovering, classifying and categorizing organisms into groups. Names are given to species when they are recognized for ecology, potential hazards or just for human culture and admiration. However, current problems such as the high diversity of species, the hundreds of publications in which species are described and the lack of a single authoritative world register routine for the definition of the taxa cause frequent confusion in the taxonomy identification. In this paper, we propose a method to automatically extract, integrate and represent species taxonomy information from the main relevant biodiversity catalogues for its comparison and studio. Consequently, the time for information analysis is reduced and taxonomic nomenclature errors are avoided, which are essential for some biology areas of knowledge.

Indexing terms/Keywords

Information Retrieval; Biodiversity Data Providers; Species taxonomy.

Academic Discipline And Sub-Disciplines

Computer science and bioinformatics.

SUBJECT CLASSIFICATION

Biology Data Managment.

TYPE (METHOD/APPROACH)

Experimental Analysis.

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 12, No.5
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

Today, the survival of wildlife is threatened with mass extinction due to loss of habitat, climate change, the introduction of non-native species and overhunting. The science of taxonomy, consisting of the discovering, naming and classification of species, is an essential task to mitigate the effects.

Detailed information of species can be queried through multiple databases [1-2]. Usually, when a scientist needs to know the taxonomy and related data about specie, it's necessary to query different databases. This work present a service to assist re-searchers searching the taxonomy, accepted name and synonyms of a specie, extracting information from biological catalogues related to life science.

The rest of this paper is structured as follows: Section 2 relates a brief description of the LifeWatch e-infrastructure. In Section 3 and 4 the last taxonomy efforts and biodiversity catalogues are described. The design and implementation of the use case are presented in Section 5 and the last section contains the conclusions drawn.

LIFEWATCH IN A NUTSHELL

LifeWatch [3] is a European research e-infrastructure (ESFRI) for biodiversity sci-ence and ecosystems research (now entering its construction phase) that involves scientists and engineers from across the European Union. It includes a range of new services and tools to help the researchers communicate, share data, analyze results, create models, manage projects and organize training.

In order to provide researchers a common point to access and share the data, the LifeWatch infrastructure includes a set of virtual labs. A virtual lab is an interoperable computing environment that allows a researcher both to update the database and to use analytical tools to extract specific information from the data. Furthermore, it also permits the collaboration with other researchers distributed across countries, time zones and disciplines.

TAXONOMIC WORK

Today, there is no automatic registration system or online inventory of all accepted species names. Consequently, animal and plant species can be named in any print publication and no mandatory register of names exists. There are no set rules governing the definition of taxa, but the naming and publication of new taxa is governed by sets of rules. Choosing the correct name is governed by international codes, the International Code of Zoological Nomenclature, the International Code of Nomenclature for algae, fungi, and plants and the International Code of Nomenclature of Bacteria.

In this context, many initiatives have been developed to facilitate the access of the taxonomy data offering visualization and analysis techniques.

USE CASE

In this section, a service for taxonomy data retrieval integrated into the LifeWatch e-infrastructure is presented. This tool facilitates the exploration of taxonomy, accepted name and synonyms of species using the information from biodiversity catalogues.

Design

The presented service is integrated in the LifeWatch architecture covering aspects related to data integration and data wrappers, as shown in Figure 1. The application permits the query of specie information in four megascience platforms: Catalogue of Life, GBIF (data portal), Encyclopedia of Life and WoRMS. All of them provide web services to access the data through external requests. To make responses of the cata-logues interoperable (which have different structure), a metadata standard for biodi-versity data called Darwin Core v.2 is used. This standard, primarily based on taxa, include a glossary of terms to provide the information exchange about biological diversity. The number of the concepts in Darwin Core is limited, but enough to repre-sent the information that the service needs to manage.

The application shows a user-friendly interface integrated in the LifeWatch website to enable a common point to access to this and other services of the infrastructure.

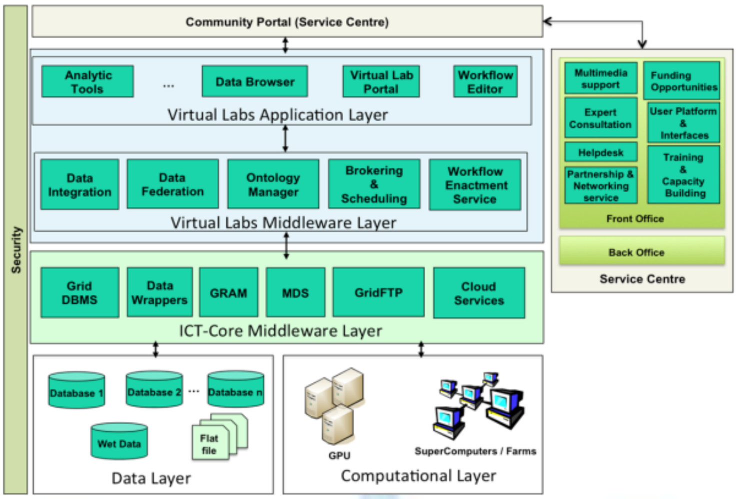

Figure 1: LifeWatch architecture in layers.

Prototype

For the implementation of the service we used a set of tools provided by the Pentaho Business Intelligence solution. First, the Pentaho Data Integration tool enables extraction, transformation and loading (ETL) capabilities to do the request of each web service, parser the response and locate the information. Then, the integration of the requests of all catalogues in single file is carried out following the schema defined by the standard Darwin Core in XML format. The result of executing the designed transformation is a XML file with the fields of the standard and the values of each database. After data extraction, the XML output file is dynamically associated with a report previously created with Pentaho Report Designer. The result is a report in HTML to be integrated in the web application.

From an implementation point of view, the interaction with the Pentaho Data Integration tool is done by the Pentaho Data Integration API, and the interaction with the Pentaho Reporting tool is done by the Pentaho Reporting Engine SDK. The process to integrate and show the result is as follows: the user introduces the required specie name and the application starts the data integration process looking for the details about the specie in all the catalogues. The generated XML with the data is passed as input to fill the report that is showed as final result in the user-friendly interface. The web application is built as a portlet to be integrated into the LifeWatch website based on Liferay which gives a single sign-on access to the application and other LifeWatch services.

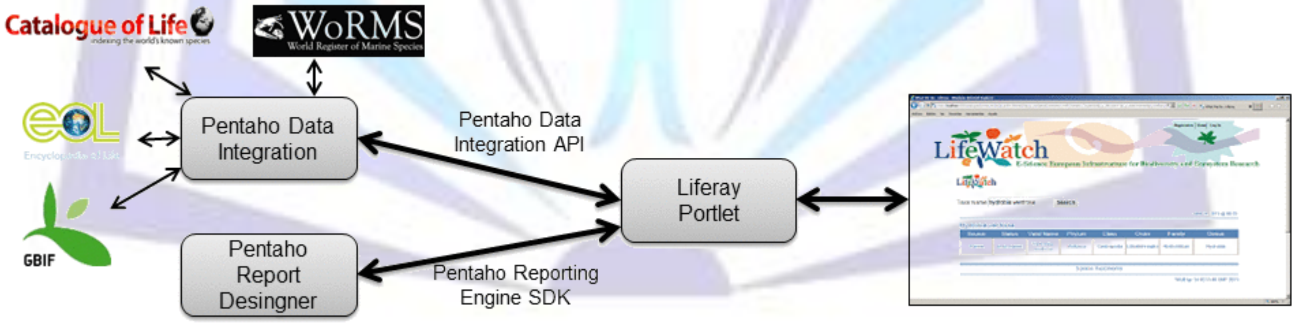
Catalogue of Life
indexing the world's known species
WoRMS
World Register of Marine Species
eol
Encyclopedia of Life
GBIF
Pentaho Data Integration
Pentaho Report Designger
Pentaho Data Integration API
Pentaho Reporting Engine SDK
Liferay Portlet
LifeWatch

Figure 2: Schema of the service for species accepted name and synonyms.

CONCLUSION

In this paper we have presented a service for species identification integrated into the LifeWatch infrastructure. This service helps scientists to analyze and compare the taxonomy, accepted name and synonyms of species. First, the data is extracted from the main species catalogues available in the Internet. Then, the Darwin Core standard and some tools for data integration, exchange and representation are used. Finally, a user-friendly interface has been implemented to facilitate the visualization of the data.

ACKNOWLEDGMENTS

This research is supported by the LifeWatch Service Centre (Italy) and the project EIC-LW-2011-0034 of the Spanish Ministry of Economy and Competitiveness.

REFERENCES

[1] Triebel D, Hagedorn G, Rambold G (2012) An appraisal of megascience platforms for biodiversity information. MycoKeys 5: 45-63.

[2] Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, et al. (2013) Global Coordi-nation and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. PLoS ONE 8(1): e51629.

[3] LifeWatch research infrastructure. http://www.lifewatch.eu

[4] Katsanevakis S, et al. (2012) Building the European Alien Species Information Network (EASIN): a novel approach for the exploration of distributed alien species data. BioInvasions Records 1(4): 235–245.

[5] Parr CS, Guralnick R, Cellinese N, Page RDM (2012) Evolutionary informatics: unifying knowledge about the diversity of life, Trends in Ecology & Evolution, 27(2): 94-103.

Author' biography with Photo

Daniel Fuentes Brenes obtained the Ph.D. degree in Computer Science in 2013 at the Seville University in Spain. He is a researcher in the Biological Station of Doñana at the Spanish National Research Council since 2012. His research interests are information systems, biodiversity collections and mobile devices. He has papers published in reputed international journals.

Nicola Fiore obtained the Ph.D. degree in Computer Science in 2004 at the University of Lecce in Italy. He is a researcher in the Ecology Unit of the Department of Biological Environmental Science and Techonologies at the University of Salento since 2009. His research interests are conceptual modelling and prototyping of Ubiquitous Applications, Data Interoperability and Big Data. He has papers published in reputed international journals.