ONTOLOGY DRIVEN FISH DATA STORAGE AND MANIPULATION

MOHD NAJIB BIN MOHD ALI

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

ONTOLOGY DRIVEN FISH DATA STORAGE AND MANIPULATION

MOHD NAJIB BIN MOHD ALI

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Mohd Najib Bin Mohd Ali

Registration/Matric No: SGR140026

Name of Degree: Master of Science (except Mathematics & Science Philosophy)

Title of Dissertation ("this Work"): Ontology Driven Fish Data Storage and Manipulation

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Ontology is a vocabulary that defines the concepts and relationships (also referred as "terms") used to describe and represent an area of concern. It is used for classifying terms of any domain of interest, which in turn characterizes possible relationships, and defines possible constraints related to the terms. Ontology provides meaning to human and computers where each ontology term will have associated metadata allowing it to have annotations, hierarchy, and relationship. Studying the role of ontologies and how to manipulate them is essential to evaluate their contribution in Semantic Web applications such as data integrations and semantic annotations. There are a number of existing fish and fisheries related databases on the internet but there are presently no specific ontology created for the fish domain. Thus there is a need to create the necessary ontology for this domain so that in the future, data for fish and fisheries can be integrated to create a large network of information. This study aims to apply semantic web applications to fish and fisheries data and to show that such data can be properly manipulated using ontology. In this study a Fish Ontology (FO) is created to show how an ontology for fish can be used to gather more information from established ontology domains related to fish, such as genetic makeup, locations, and diseases. The Fish Ontology in this study demonstrates the possibility of using ontology as an automatic fish classification tool. The methods presented in this study enable automated classification of a fish specimen based on its taxon rank, using the FO, showing how data within the ontology can be linked to other data using data manipulation such as data extraction, or deletion. Future studies should include more species in the ontology model, improved annotations, and more revised terms.

ABSTRAK

Ontologi adalah kosa kata yang menentukan konsep dan hubungan (juga dirujuk sebagai "istilah") digunakan untuk menggambarkan dan mewakili sesuatu domain. Ia digunakan untuk mengklasifikasikan istilah domain yang diminati dengan mencirikan kemungkinan untuk setiap hubungan, dan menentukan kemungkinan untuk setiap kekangan yang berkaitan dengan istilah tersebut. Ontologi memberi makna kepada manusia dan komputer di mana setiap istilah didalam ontologi mempunyai metadata, membenarkan istilah tersebut mempunyai anotasi, hierarki, dan hubungan. Mengkaji peranan ontologi dan cara memanipulasikannya penting untuk menilai sumbangannya terhadap aplikasi Web Semantik seperti integrasi data dan penjelasan semantik. Terdapat banyak pangkalan data sedia ada berkaitan dengan ikan dan perikanan di internet, namun pada masa ini tiada lagi ontologi yang khusus dicipta untuk domain ikan. Oleh itu terdapat keperluan menciptanya supaya kelak, data tersebut boleh digabungkan untuk mewujudkan rangkaian maklumat yang luas. Kajian ini bertujuan untuk mengaplikasikan web semantik terhadap data ikan dan perikanan, dan mampamerkan bahawa data tersebut boleh dimanipulasikan menggunakan ontologi. Di dalam kajian ini “Fish Ontology” (FO) dicipta untuk menunjukkan kebolehan ontologi ikan mengumpul maklumat daripada domain lain yang berkaitan, seperti genetik, lokasi, dan penyakit. “Fish Ontology” di dalam kajian ini menunjukkan kemungkinan menggunakan ontologi sebagai alat pengklasifikasian ikan secara automatik. Kaedah yang dibentangkan dalam kajian ini membolehkan pengkelasan spesimen ikan secara automatik berdasarkan pangkat takson, menggunakan FO, menunjukkan bagaimana data didalam sesebuah ontologi boleh dikaitkan dengan data-data yang lain melalui kaedah manipulasi data seperti pengekstrakan dan pemadaman data. Kajian di masa hadapan haruslah merangkumi lebih banyak spesies untuk model ontologi yang sedia ada, berserta dengan anotasi data yang lebih baik, dan istilah yang disemak semula.

ACKNOWLEDGEMENTS

I would like to extend my appreciation and gratitude to all who have helped me to complete this research. Their commitment to oversee the completion of this study is crucial and has been beneficial in every aspect of the project development.

I would like to thank my supervisor, Associate Professor Dr. Sarinder Kaur Kashmir Singh, for her ideas which led me to start this ontology based project, her advice which always points me to the proper path, her support, and encouragement on seeing the completion of this study.

I would also like to thank my co-supervisor, Dr. Amy Then Yee Hui, for all of her assistance and guidance on the study. She has been really helpful by giving proper advice on fish-related terms for the ontology and on the structure of the ontology. She also provided the book and paper necessary for the completion of this study. Her strict advice, careful reminder and continuous encouragement have helped me to ensure that the study is completed accordingly.

I would also like to thank Professor Dr. Chong Ving Ching for providing the necessary data for this study. He has been providing fish sampling and diversity data that are relevant to this study and provided clues to creating terms for the ontology. The data provided are part of his research conducted years ago, kept on magnetic disk and papers, and I am honored to take part in converting some of these data to digital form for safekeeping in hard disks.

Finally, I would like to thank fellow lab members such as Miss Aqilah, Miss Elham, Mr. Haris Ali Khan, Mr. Liow Lee Kien, Mr. Teo Bee Guan, Mr. Khoo Soon Jye, some of my colleagues such as Mr. Ahmad Fadel Berakdar, and Mr. Ubaid Ur Rehman, and the staff of Bioinformatics building such as Miss Sugunadevi Rajagopal, Mr. Kamaruddin, and Mr. Ridzuan for all of their help and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

API	:	Application Program Interface
BMP	:	Bitmap Image File
CEC	:	Commission of the European Communities
CRM	:	Customer Relationship Management
CSV	:	Comma-Separated Values
DB	:	Database
DL	:	Description Logic
FAO	:	Food and Agriculture Organization of the United Nations
FASTA	:	Fast Alignment Search Tool – All.
FISHBOL	:	Fish Barcoding Of Life
FO	:	Fish Ontology
FOAF	:	Friend of a Friend
FOS	:	Fishery Ontology Service. Fisheries Ontology of FAO.
GO	:	Gene Ontology
GUI	:	Graphical User Interface
ICLARM	:	International Center for Living Aquatic Resources Management
IUCN	:	International Union for Conservation of Nature
KAON	:	Karlsruhe Ontology
LSID	:	Life Science Identifiers
MHBO	:	Monogenean Haptoral Bar Image Ontology
NCBI	:	National Center for Biotechnology Information
NIWA	:	NZ Freshwater Fish Database
OBO	:	Open Biomedical Ontologies
OOPS	:	Ontology Pitfall Scanner Tool

OWL	:	Web Ontology Language
PDF	:	Portable Document Format
RDF	:	Resource Description Framework
RDF4J	:	Ontology Portal Framework known as SESAME
RDFS	:	Resource Description Framework Schema
RSS	:	Rich Site Summary
SAIL	:	Storage and Inference Layer
SeRQL	:	Second Generation RDF Query Language
SWRL	:	Semantic Web Rule Language
SQWRL	:	Semantic Query-Enhanced Web Rule Language
SESAME	:	Ontology Portal Framework known as RDF4J
SPARQL	:	Simple Protocol and RDF Query Language
SQL	:	Structured Query Language
TDWG	:	Taxonomic Database Working Group
TLO	:	Top Layer Ontology
TTO	:	Teleost Taxonomy Ontology
TXT	:	Filename extension for text files
URI	:	Uniform Resource Identifier
URL	:	Uniform Resource Locator
VTO	:	Vertebrate Taxonomy Ontology
VSAO	:	Vertebrate Skeletal Anatomy Ontology
WWW	:	World Wide Web
XLS	:	Microsoft Excel file format
XML	:	Extensible Markup Language
XSD	:	XML Schema Definition
ZFIN	:	The Zebrafish Model Organism Database

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Ontology, one of the most important aspects in semantic web applications, has become an indispensable tool in the field of data management. It plays a significant role in biodiversity and biomedical research as an underlying framework and architecture of a variety of applications. Semantic Web is the next generation of World Wide Web, an extension of the current web which enable computers and people to work in cooperation. Ontology on the other hand is the vocabulary that defines the concept and relationships of any area of concerns which are used by the semantic web applications. Ontology is one of the most fundamental components of semantic web (Berners-Lee et al., 2001), and is primarily used as a source of vocabulary for standardization and integration purposes. Additionally, some applications use ontologies as a basis of computable knowledge. (Bollier & Firestone, 2010). The semantic web technology provides a promising platform for biodiversity researchers to link and share data, in order to integrate information using the World Wide Web (Deans et al., 2012).

With the exponential growth of biodiversity data, it would be beneficial to restructure current datasets into formats compatible with the semantic web applications and technology. This development would be best achieved by the collaboration of domain experts and ontology specialist. An ontology that is created for a domain will make the data and terms for that domain more meaningful for human understanding and more optimized for computers consumption to achieve more intelligent applications (Page, 2006). Biodiversity data like fish datasets are usually stored using relational database model, focusing on species related information (Alroy et al., 2012; Frimpong & Angermeier, 2009; Froese & Pauly, 2017; Great Lakes Fishery Commission, 2009; Ickes et al., 2003; International Game Fish Association, 2015; Nelson, 2006; NIWA, 2016; Shao, 2001; Ward et al., 2009). Data in these repositories are usually structured

based on the researcher's interests and needs, which restricts the generation of uniform naming standards. Hence, ontologies can facilitate this by generating structured vocabularies that describe entities of a domain of interest and their relationships with each other (Shadbolt et al., 2006). Species information generated by an ontology will likely be more optimized for human readability and will lay the underlying foundation upon which applications can be integrated with each other.

1.1 Overview

Fish data can be found in abundance and scattered around the web. Most of these data are stored in a variety of forms, having different meaning depending on the interest of the data curator. Species morphology description, genetic makeup, fish anatomy, habitat distribution, and publication content are some of the accessible data of interest to most of the scientific community working on fish and fisheries research. Most of these datasets usually need to be simplified or cleaned before being made available online for ease of human understanding; however, some data are very complex and can only be analyzed efficiently with the help of specific computer programs. Catch records, individual specimen details, and biomass distribution are some examples of data that hold a lot of raw information. They can be too large to be uploaded on the web and are difficult to be interpreted by humans. On occasions, when converting the raw datasets to be published online, lots of potentially useful data is lost in the cleaning process. This loss of data can likely be eliminated or reduced by the application of standardized vocabularies for the generation of integrated applications.

Large raw data usually have a wide range of information, such as image attachments, genetic marker information, and hereditary information. Sometimes there are unused information attached such as unit number, sample size or date of catch. Wide data type such as table, text, graph, genetic coding, and image generate a variety of data

formats and extensions such as XLS, SQL, TXT, FASTA, PDF, and BMP. Usually, there is no clear way to merge these wide ranges of data formats. The usage of ontology and semantic web technology, however, makes it possible to integrate the different data sets and format types together, assisting data analysis application.

Assembling the data sets needed for global biodiversity needs has always been challenging. There are about 2 to 3 billion specimens estimated to be in the world's biological collection, however, only less than 10% have been recorded in databases and digital images (Ariño, 2010; Duckworth et al., 1993). Biodiversity data such as information about organisms, morphology, genetics, life history, habitats, and geographical distribution are highly heterogeneous. These datasets usually contain spatial, temporal, and environmental data. Biodiversity science seeks to understand the origin, drives, and function of this variation, thus requires integrated data on the spatiotemporal dynamics of organisms, populations, and species, together with information on their ecological and environmental context. Since biodiversity knowledge is generated across multiple disciplines, each with its own community practices, most of the data are stored in a fragmented network of resource silos, in formats that hinder integration. In order for these sources to fulfill their potential in terms of flexibility, usage and re-usage in a wider variety of monitoring, scientific, and policy-oriented applications, it is essential to find the means to properly describe and interrelate the data types and sources (Hardisty et al., 2013).

The need to standardize biodiversity vocabulary is not recent. Ontology is the vocabulary which defines the concepts and relationships (also referred as "terms") within an area of concern. It is used for classifying terms within a domain of interest, characterizing possible relationships, and defining possible constraints related to the

terms. The role of vocabularies on the semantic web is to help data integration when, for example, ambiguities may exist on the terms used in the different data sets, or when additional knowledge may lead to the discovery of new relationships. This is due to its capabilities to handle big data and linked data application. Ontologies extract relevant data from a source application, such as a Customer Relationship Management (CRM) system, big data applications, files, warranty documents, etc. These extracted data or semantics are linked into a search graph instead of a schema to retrieve results, enabling users to search a schematic model of all the datasets that are linked to each other within the network of integrated set of applications (Lanace, 2014).

In the past years, many enterprise applications have been developed and used by organizations for various needs and with various requirements. Integrating applications to obtain a company-wide integrated view is difficult, expensive and often not without risks. Ontology introduces a new way to use enterprise applications. It allows users to search, link and integrate their applications, databases, files, and spreadsheets anywhere. Ontology eliminates the need to integrate systems and applications when looking for critical data or trends since it uses a unique combination of an inherently agile, graph-based semantic model and semantic search to reduce the timescale and cost of complex data integration challenges.

Fish can be described as any non-tetrapod chordate (four footed animals), that has gills throughout life and has limbs, if any, in the shape of fins (Nelson, 2006). Data generated from fishing and fisheries activities, in addition to species-specific information, are huge. Most of them are related to sampling, genetic and taxonomic data. This huge datasets are obvious given that the total number of fish species has been estimated at 32,000 to 40,000 globally (Nelson, 2006). Various data such as location,

morphology, species information and population can be gathered for any fish species. Usually, these data types, if made available by the owner, are scattered around the web. A centralized storage location to store the data for most of these different data types and sources will allow better data management and linkage. Data and knowledge can be linked together and can be managed better with the help of ontology which is one of the main driving force for the new version of the web (Chang & Terpenny, 2009). Since ontology has the potential to drive data acquisition, correlation and migration projects in a post-Google world, it is perfect to be used as the base for this research.

1.2 Research Question

This study aims to answer the following research questions:

(1) What are the available databases or computer systems that cover the topic on fish in the public domain?

(2) What are the terms used to represent the data contained in these fish-related systems?

(3) Are these systems integrated and what are the options available to integrate data?

(4) What is the best solution in managing fish-related data that is in line with the current technology and trends?

1.3 Research Objectives

This study aims to explore the application of ontology and semantic web applications in the biodiversity domain, fish in particular. The objectives of the study are:

(1) To improve current fish biodiversity data representation using ontology and semantic web.

(2) To propose a standard vocabulary in the fish and fishery domains.

(3) To propose a solution for a standardized and comprehensive fish-related ontology that can facilitate data integration in the fish and fishery domains.

1.4 Research Approach

To achieve the first objective, 11 published online ontologies, 4 terms standard and 3 real life applications (Table 1.1) were observed and studied in order to fully grasp the capability and potential of ontology and semantic web application. Some of the most important ones are selected and discussed in the results section (Table 3.2).

Table 1.1: Popular terminologies observed from databases, ontologies and books.

Sources	Description
TDWG LSID	Vocabularies or descriptions of the metadata returned for particular classes of object within the TDWG domain. Form part of a larger TDWG ontology effort that describes how these classes of data are related. Can be used in any XML or Semantic Web based technology to express concepts associated with biodiversity.
OBO Foundry	Collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate.
The Diversity of Fishes: Biology, Evolution, and Ecology 2nd Edition	Books that represents a major revision of the world's most widely adopted ichthyology textbook. The text incorporates the latest advances in the biology of fishes, covering taxonomy, anatomy, physiology, biogeography, ecology, and behavior.
Shark and Rays of Borneo	Books that are the first comprehensive reference on the sharks and rays of Borneo. It is the result of a collaborative project between the governments of the United States, Malaysia, Indonesia and Australia, and is funded by the National Science Foundation.
Gene Ontology	An ontology that provides controlled vocabularies of defined terms representing gene product properties. These cover three domains: Cellular Component, the parts of a cell or its extracellular environment; Molecular Function, the elemental

Table 1.1: continued.

	activities of a gene product at the molecular level, such as binding or catalysis; and Biological Process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units
Vertebrate Taxonomy Ontology (VTO)	An ontology on vertebrate taxonomy which includes both extinct and extant vertebrates. Its hierarchy backbone for extant taxa is based on the NCBI taxonomy complemented by taxonomic information across the vertebrates from the Paleobiology Database (PaleoDB), the Teleost Taxonomy Ontology (TTO) and AmphibiaWeb (AWeb) to provide a more authoritative hierarchy and a richer set of names for specific taxonomic groups.
Disease Ontology	An ontology that been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts
Zebrafish Anatomy Ontology (ZFO)	A structured controlled vocabulary of the anatomy and development of the Zebrafish (*Danio rerio*).
Chemical Entities of Biological Interest Ontology (ChEBI)	Ontology of a freely available dictionary for molecular entities focused on 'small' chemical compounds. It incorporates an ontological classification, and uses nomenclature, symbolism and terminology endorsed by the 2 international scientific bodies which are the International Union of Pure and Applied Chemistry (IUPAC) and the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)
Epidemiology Ontology (EPO)	An ontology which are designed to support the semantic annotation of epidemiology resources. It is being developed under the EU-funded EPIWORK project, a multidisciplinary research effort which aims at increasing the amount of epidemiological data available, improving disease surveillance systems, and promoting the collaboration among epidemiological researchers.
Teleost Taxonomy Ontology (TTO)	An ontology covering the taxonomy of teleosts (bony fish) which is being used to facilitate annotation of its phenotypes, particularly for taxa that are not covered by NCBI. It serves as the source of taxa for identifying evolutionary changes that match the phenotype of a zebrafish mutant.
Pizza Ontology	An example ontology that contains all constructs required for the various versions of the Pizza Tutorial run by Manchester University.

Table 1.1: continued.

Marine Top Layer Ontology (MarineTLO)	A Top Level Ontology for the Marine Domain. It is the Conceptual backbone of the MarineTLO-based warehouse, which integrates information coming from FishBase, WoRMS, ECOSCOPE, FLOD and DBpedia. It currently contains information of around 3M triples about marine species and 40,000 ecosystems, water areas, vessels, etc. The warehouse is already in use by various services offered by iMarine.
Common Anatomy Reference Ontology (CARO)	An upper level ontology to facilitate interoperability between existing anatomy ontologies for different species. It is being developed to facilitate interoperability between existing anatomy ontologies for different species, and will provide a template for building new anatomy ontologies.
NCBI organismal classification	An ontology representation of the NCBI organismal taxonomy which would automatic translate the datasets of the NCBI taxonomy database into obo/owl.
NCBITaxon	An online database which is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.
FishBase	An online relational database with information to cater to different professionals such as research scientists, fisheries managers, and zoologists. It contains 3300 fish Species, 318500 Common names, 57400 Pictures, 53000 References, and have 2250 Collaborators which works on the database.
PaleoDB	An online relational database for paleontological data which has been organized and operated by a multi-disciplinary, multi-institutional, international group of paleobiological researchers. Its purpose is to provide global, collection-based occurrence and taxonomic data for organisms of all geological ages, as well data services to allow easy access to data for independent development of analytical tools, visualization software, and applications of all types. The Database's broader goal is to encourage and enable data-driven collaborative efforts that address large-scale paleobiological questions.

To achieve the second objective, an ontology is created based on sample data as well as by referring to popular ontologies. Sample data is cleaned and reviewed by domain experts before it is used in this study.

To achieve the last objective, the work on the ontology is published to ensure that the structure is agreed upon by experts. Furthermore the ontology is reviewed by fish experts to validate the usefulness of its application.

1.5 Outline of the study

Chapter One: This chapter outlines the need for using ontology, which is the key element in the semantic web application. The introduction section explains the need of using ontology, and the need to change the current fish data set environment, besides presenting the research questions, objectives and approach of this study.

Chapter Two: This chapter contains the literature review, which provides background about the best way to handle data on the web, and ontology versus popular database environment. This chapter also explains about ontology structures, practices, tools, framework, developing environment and portal and provides good ontology example. Some background information about the related studies is also included in this chapter.

Chapter Three: This chapter contains the methods and materials used to create the ontology, the portal, and evaluation. The methodological flow is presented firstly, followed by details on data acquisition, and ontology creation. Later, the term addition is being elaborated, and finally, the chapter is ended by explaining the method to evaluate the ontology.

Chapter Four: This chapter presents the results of the created ontology framework, its relationships, integration with other sources, inferencing capabilities, and querying capabilities. Also presented in this chapter is the results of the portal created specifically for this ontology, its framework, and capabilities, and lastly, the results from evaluating the ontology.

Chapter Five: This chapter discusses the results obtained in the ontology and portal creation. It also contains comparisons for sources that can be included in the ontology, further explaining its features and the reason why it is or not being included in the ontology. Furthermore, this chapter also discusses the issues encountered in the course of the studies, revolving around the ontology coverage, terms importance, tools, evaluations, and semantic web applications. Later discussed in this chapter are the strengths and weaknesses of the ontology created in this study, its evolutions and future directions, declaration on the future enhancement of the ontology model, and finally conclusions.

CHAPTER 2: LITERATURE REVIEW

In the world of semantic web, linked data and big data can be described as the building blocks of the next generation web, ensuring the evolution of data from the web 2.0 (user-generated content) to web 3.0 (semantic web). There are five criteria in order for data to achieve a 5 star rating, namely, (1) data of any format should be available on the Web under an open license, (2) data should be available as structured data (e.g., Excel instead of image scan of a table), (3) data should be available in a non-proprietary open format (e.g., .CSV as or .XLS), (4) URIs should be used to denote things, so that the designated data can be pointed, and (5) data should be linked so that exact data are connected to other data providing context (Berners-Lee, 2009; Berners-Lee et al., 2015). Most of the web 2.0 data only have achieved 3 to 4 star criteria. The fifth one, which is to ensure that data are linked together, is usually neglected but it is one of the most important components which enable the dataset to evolve from web 2.0 to web 3.0.

To prepare data for semantic web, the creation of an ontology is crucial since an ontology can define the naming, types, properties and relationships of any terms which exist in the domain coverage (Chang & Terpenny, 2009). Currently, there are several important ontology structures prepared by several groups who are enthusiastic on the development of semantic web technology. The Web Ontology Language (OWL) Working Group (W3C OWL Working Group, 2009) and the Open Biomedical Ontologies (OBO) Foundries (Smith et al., 2007) are some of the most important groups involved in ontology project. Although there are considerable difference between their format structures (OWL and OBO), both are known to provide ontology guidelines in handling big data and providing metadata capabilities to the created ontology (Golbreich et al., 2007; Tirmizi et al., 2011).

While there are debates on which of the two is better suited for creating ontology, the choice would likely be based on the user's needs. There are claims that scientists prefer the use of the OBO file format while data engineers would like to use the OWL file format. The OWL file format focuses more on automatic reasoning using logic while the OBO format focuses on supporting existing users. Hence the background for both of these file formats differs as well where the OWL format favors more to Artificial Intelligence, which is preferred by the data engineers while the OBO format favors more to terms annotations which are favored by the scientists. As such, the usage differs where the OWL format describes any domain in theory due to its generic approach (top-down) while the OBO format which is used mainly by biologist, describes biology in practice since it is more specific (bottom-up). As example, in OBO, you need to define "name: leg", and "relationship: part_of thoracic segment", while in OWL you can write it as "leg *SubClassOf* part_of some thoracic segment". However, in the recent years, there is a lot of ontological work in science that provides both files format to represent their work. Since there are some similarities between the two, we finally agreed to use the OWL file format while following the guidelines set by the OBO Foundry. In this way, the created ontology will be able to relate to both of the file formats, allowing easy future integration and communication to any related ontology to fish domain (Smith et al., 2007).

To create an ontology, several steps or precautions must be followed. These include (1) determining the domain and scope of the ontology, (2) considering to reuse existing ontologies, (3) enumerating important terms in the ontology, (4) defining the classes and the class hierarchy, (5) defining the properties of classes, (6) defining the facets of the slots, and (7) creating instances (Noy & McGuinness, 2001). These steps ensure that the created ontology are well structured, maintained, and linkable to other data related to its domain, thus are followed in this research.

As the semantic web research advances, there are a number of tools that can aid ontology creation. Altova (Altova, 2016), NeoN Toolkit (Neon Foundation, 2016), TopBraid Composer (Top Quadrant, 2016), KAON (Motik, 2005), and Protégé (Protégé, 2016) are some of the most popular online tools. Ontology editors and tools usually vary according to the purpose of the project and the kind of file format it can support. Some are created as a programmable XML editors used for knowledge extraction which transforms Web pages into RDF format, some works as a visual RDF and OWL editor that automatically generates RDF/XML files or nTriples files (both are common formats for semantic web development aside from OWL and OBO file format) based on visual ontology design, and some work as a vocabulary prompting tool to help assist human in managing its vocabulary resources. Regardless of the purpose these tools are created for, either it is for ontology editing, ontology mapping, or ontology visualization and analysis, it is imperative to find proper tools which suit the need of the developer to ensure the created ontology is well built and thoroughly developed.

A good ontology creation tool must be able to provide various feature to ensure that it is easy for the user to view the ontology structure, import and export terms, view all the terms and metadata, link and integrate terms, and have the capability to standardize the data and metadata. Protégé is one of the software that provides these features since it has many supporting tools which can help users in creating their own ontology. Besides, it is free, open source, has a user-friendly GUI, and it supports the new Ontology Web Language formats such as OWL (Bechhofer, 2009; W3C OWL Working Group, 2009) and OWL2 (W3C OWL Working Group, 2012). It comes with important built-in plugins useful for complete ontology development. Protégé also supports the ontology reasoning plugins, visualization plugins, and ontology querying plugins. There are also some external plugin that can be downloaded that can help users to build a solid ontology.

There are several requirements for creating a knowledge base on which future simulation can be built upon, while ensuring their semantic coherence and operational interoperability. An ontology must be able to handle unstructured information as input sources, reusing existing knowledge base and information, must be able to handle formal and informal representation, data and terms must be credible, verifiable, authentic, consistent, and validated. It also must allow quick and easy development (understandable and easy to use terms and structure), action-centric (not focusing on concept, but rather real life application), and lastly it also must be flexible and adaptable (Doumeingts et al., 2007).

Available standards and guidelines can be followed to create a useful ontology. For example, Taxonomic Database Working Group Life Science Identifier (TDWG LSID) (Orme et al., 2008) and Darwin Core (Wieczorek et al., 2012) contain terms which are also relevant in the fish domain. However, the usage of both of these standards has been quite slow recently due to data integration issues. In 2007, the successful creation of Gene Ontology (Ashburner et al., 2000) gave birth to an organization known as OBO foundries (Smith et al., 2007), which started an initiative in medical science domain with several guidelines to create an ontology which is interoperable, logically well-formed, and to incorporate an accurate representation of biology reality. The approach taken by this organization is widely accepted, and currently there are around 150 ontologies followed their guidelines.

Standards aside, ontology validation is also one of the most important aspects that must not be overlooked when creating an ontology. Data and terms that have been incorporated in the ontology must be validated either manually or automatically with the help of computer inferring capabilities to ensure the integrity of the ontology. The logical representation of the terms and its relationships must allow inference engines to

test for semantic interoperability (Glimm et al., 2014; Sirin et al., 2007; Tsarkov & Horrocks, 2006). Aspects that are usually checked for ontology validation are mostly on content validation (evaluate individual messages given the axiom of the reference ontology), information flow validation (determine that the message is being sent and received in an appropriate order), process flow validation (determine whether the event captured by the terms and relationship in the ontology meet the requirements of process model), consistency validation (determine whether the available information is consistent within and across the messages), and assertion validation (using additional or external knowledge to evaluate information) (Kalfoglou, 2009).

Semantic web framework is also another important aspect in ontology creation. It classifies the different Semantic Web technologies according to their functionalities and represents them as independent components, providing description of their functionalities, and provides dependencies between the components (García-Castro et al., 2008). Apache Jena is an open source Semantic Web framework for Java (Apache Jena, 2016). It provides an Application Program Interface (API) to extract data from and write to the Resource Description Framework (RDF) graphs which are the underlying structure of ontology. These graphs are represented as an abstract "model" integrating data from files, databases, URLs or a combination of these.

Apart from Apache Jena, Eclipse RDF4J (formerly known as Sesame) is a powerful Java framework alternative for processing and handling RDF data (Eclipse RDF4J, 2016). This includes creating, parsing, scalable storage, reasoning and querying with RDF and Linked Data. It offers an easy-to-use API that can be connected to all leading RDF database solutions. Being governed by the Eclipse Foundation means a stable, vendor-neutral steward takes responsibility for continued support of the RDF4J project. Eclipse’s rigorous IP review and quality control structures give users of RDF4J the

assurances they need for safe use of the framework in enterprise environments. Eclipse being a very recognized and trusted brand with a large open source community will help RDF4J attract more users and developers, ensuring its long-term growth and development.

The last element that complements the ontology development is the semantic web portal. A web portal is defined as a collection of relevant links to text, voice, video image, emails or other relevant data on a single Web page (Sathyanarayan, 2004). A semantic portal in the other hand is a web portal which is built based on W3C Semantic web standards, where it differs from the traditional design in several ways, such as it can support multidimensional search capabilities with the help of rich domain ontologies, with semi-structured and extensible information which allows for bottom-up evolution and decentralized updates (Reynolds & Shabajee, 2001).

Ontologies can represent many domains of knowledge whilst being machine understandable. However, traversing large ontologies and fulfilling specific user demands, often takes many computing hours to complete.

2.1 Related Studies

There is an abundance of fish data scattered around the web in the form of web portal and databases, and many ontologies have been created for biodiversity (Abu et al., 2013; Avraham et al., 2008; Caracciolo, 2007; Dahdul et al., 2010; Federhen, 2016; Gangemi et al., 2004; Midford et al., 2010, 2013; Seltmann et al., 2012; Sprague et al., 2003; Tzitzikas et al., 2013, 2016; Van Slyke et al., 2014; Yoder et al., 2010; Zheng et al., 2010). Most of the databases or web portals show different kind of fish data published by the web authors to share their information and findings with the public according to their specialty and interest (Frimpong & Angermeier, 2009; Froese &

Pauly, 2000, 2017; Great Lakes Fishery Commission, 2009; Ickes et al., 2003; International Game Fish Association, 2015; NIWA, 2016; Shao, 2001; Ward et al., 2009). Most of the public data available are concerned more about species details, taxonomic information, habitat, and genetic information.

2.1.1 Fish Databases

In 1991, the International Center for Living Aquatic Resources Management (ICLARM) in collaboration with the Food and Agriculture Organization of the United Nations (FAO) and with the support of the Commission of the European Communities (CEC) developed the FishBase (Froese & Pauly, 2000, 2017) to summarize global information on finfish. This database contains the most comprehensive information about fishes, from contributors all around the world.

In 2001 another version of the fish database which covers the fishes in Taiwan emerged (Shao, 2001). This database, called the "The Fish Database of Taiwan", complements the FishBase. It has information on fish hierarchy, taxonomy, distribution, specimen, and reference for fishes found in Taiwan. The fisheries scientists would use both websites to fully confirm the information about a fish species, especially if the species can be found in Taiwan.

2.1.2 Gene Ontology

In 2000, the Gene Ontology (GO) was constructed to document information about genes. The project, created as a 3 layered domain information structure, contains information on gene biological process, gene cellular component, and gene molecular function (Ashburner et al., 2000). The GO database integrates the vocabularies and contributed annotations and provides full access to this information in several formats.

Members of the GO Consortium continually work collectively, involving outside experts as needed, to expand and update the GO vocabularies. The GO Web resource also provides access to extensive documentation about the GO project and links to applications that use GO data for functional analyses.

The gene ontology is similar to the Fish Ontology developed in this study, in terms of annotations and its unique ID formatting. In fact, the FO follows similar standard provided by the GO in order to achieve high integration value in the future.

2.1.3 Pizza Ontology

Another popular ontology which has a similar structure is Pizza Ontology developed by the Manchester University (Horridge et al., 2011), created using Protégé. This ontology provided the terminology on Pizza, and all the necessary relationships to determine a pizza. The similarity between Pizza Ontology and Fish Ontology is shown in their relationship structure which allows these ontologies to automatically infer information to determine any terms or classes relationships. Both these ontologies can automatically provide new information based on several restrictions given to them, where they can find new information on any terms.

CHAPTER 3: METHODS & MATERIALS

In this chapter, the research methodology is described in detail in the following sections: Data Acquisition and Cleaning, Ontology Creation, Portal Creation, Ontology manipulation through portal and tools, and evaluation. The approach followed the project flowchart illustrated in Figure 3.1 for ontology creation and Figure 3.2 for the prototype web portal development while Table 3.1 shows the list of tools that were used in this research.

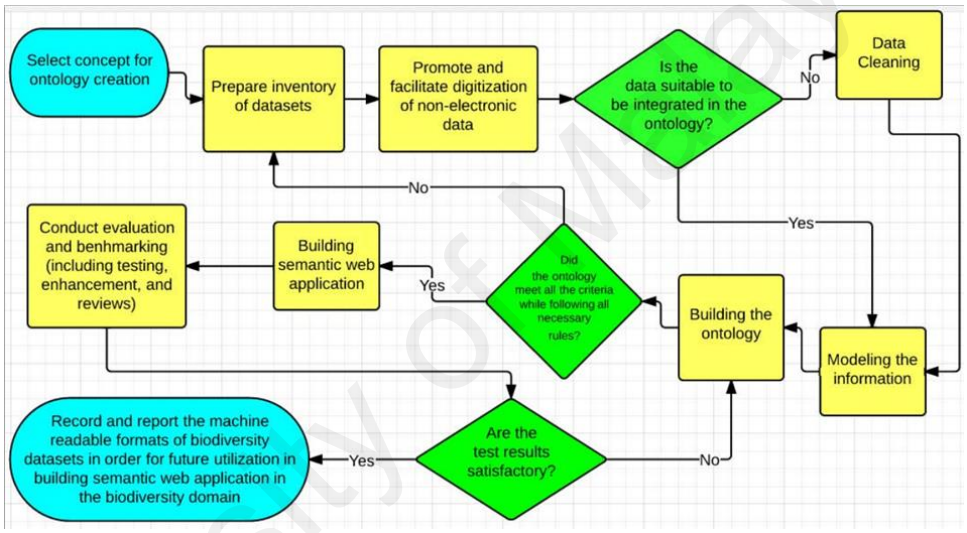


Figure 3.1: Workflow of study.

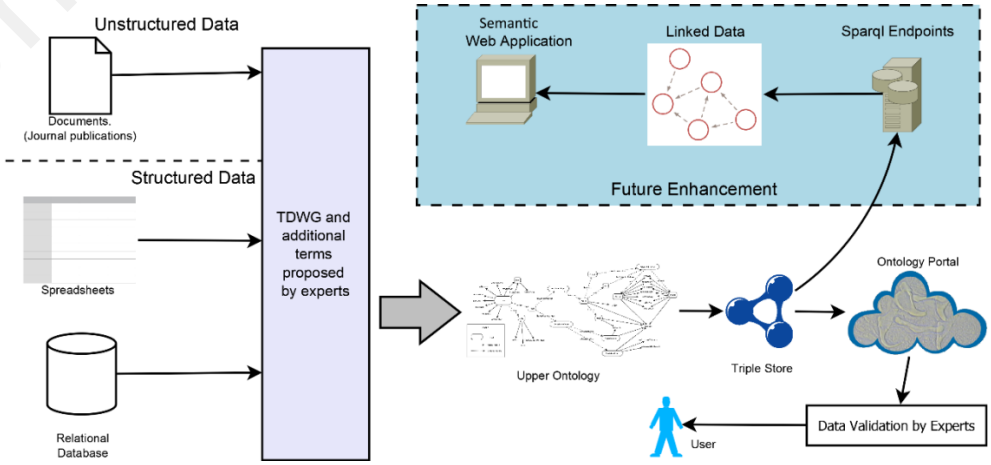


Figure 3.2: Workflow for portal development.

Table 3.1: List of tools used in the research and their functions.

Type	Name	Functions
Operating System	Microsoft Windows	Operating system for running necessary programs for the project development.
Data Analysis, Ontology Designing	Microsoft Office, Dia Diagram	Tools necessary to read and analyze data
Ontology Creation and Data Population	Protégé	Editor for ontology. Contain useful plugins such as OWLViz and Ontograf to visualize the created ontology, SPARQL query editor to test the triples query in the created ontology, and Reasoners to automatically infer the concept relationship.
Ontology Portal Creation	Apache Jena, Sesame RDF, Eclipse IDE, Netbeans IDE	Apache Jena and Sesame RDF are the framework used to connect ontology data with the portal. The portal are created as a Java Web based Applications using Eclipse or Netbeans as the IDE.

3.1 Data Source

Fish data used in this research were obtained from 2 sources which were: 1) Professor Dr. Chong Ving Ching data from 1980 to 2000 of fish from Matang Selangor, and 2) Public online databases such as FishBase (Froese & Pauly, 2000), and IUCN Red List of Threatened Species (IUCN, 2016). The fish data acquired from both sources are used to fill up the species and specimen data in the ontology and to provide metadata to each of the species. Data acquired from these sources are stored as a flat data in Microsoft Excel. The data is then further examined for its suitability to be adapted into the ontology. Subsequently, data is cleaned up to ensure that there is no error during conversion into an ontology.

3.2 Ontology Creation

Ontology creation is divided into 2 parts, which are terms and relations, and terms validation, explained in the subchapters below.

3.2.1 Terms and Relations

The terms incorporated in the Fish Ontology were based on research from the following sources: TDWG standard (LSID and Darwin Core) (Orme et al., 2008; Wieczorek et al., 2012), the book “The Diversity of Fishes” (Helfman et al., 2009), and several ontology related to this research domain (the complete list is presented in Table 3.2). The criteria adopted for selecting the terms and relationships needed in the creation of the ontology are based on several factors which are:

1. Whether the terms have already been used by other ontology.
2. Whether the terms are usually used or covered by the related domain.
3. Whether the terms have different meaning and use.
4. Whether the usage of the terms can affect the structure of the ontology.
5. Whether the terms can change the meaning and functions of the ontology.
6. Whether the source of the terms gave "free to use" permission.

The terms are taken from various sources in order to increase the granularity of the created ontology.

Table 3.2: Terms sources list.

Sources	Terms usage description
TDWG LSID (Orme et al., 2008)	Provided terms, structure and relationships for general terms (E.g.: Taxon and Location).
The Diversity of Fishes (Helfman et al., 2009)	Provided terms related to fish taxonomy rank, fish anatomy, fish history, and fish details.
Vertebrate Taxonomy Ontology (VTO) (Midford et al., 2013)	Provided terms, relationships, data and annotations for vertebrate's species. Only species related to fish are selected to minimize ontology size.
Teleost Taxonomy Ontology (TTO) (Midford et al., 2010)	Provided terms, relationships, data and annotations for teleost species including taxon rank and anatomy.
NCBITaxon (Federhen, 2016)	Provided species terms and relationships for any fish species not covered by the VTO
MarineTLO (Tzitzikas et al., 2016)	Provided terms which are related to marine species, which will help fish ontology to be integrated to upper layer ontology
FishBase (Froese & Pauly, 2000, 2017)	Provided metadata for fish. Included in the ontology as annotations link.
PaleoDB (Alroy et al., 2012)	Provided metadata for fish fossil. Included in the ontology as annotations link.

Most of the terms added to the ontology were assigned with annotations to increase the granularity of the ontology. Furthermore, most of the metadata included in the ontology mainly describes the terms description, the ID for the original terms, label, namespace, synonyms and cross-references. Table 3.3 below shows some examples of the terms in the Fish Ontology adopted from the sources (Table 3.2) in this research.

Table 3.3: Terms adoption in the Fish Ontology.

Example of Terms	**Sources**			**Implementations in the Fish Ontology**
	Helfman (2009)	Vertebrate Taxonomy Ontology (VTO)	NCBITaxon	
Furcacaudi-formes (order)	Classified as Subclass of Thelodonti (superclass)	Classified as subclass of Agnatha (class)	Not classified	Follows and reuses the VTO terms
JawlessFish	Contains species and information for jawless fish species	No classes and annotations found, but related species are classified	No classes and annotations found, but related species are classified	Follows Helfman (2009) for labeling
LobeFinned Fish	Classify it as Actinopter-ygii (page 4)	No classes and annotations found, but related species are classified	Classified as Coelacanthi-formes	Follow Helfman (2009) for classification and labeling
Gobiidae (family)	Listed and classified as family	Listed and classified as family.	Listed and classified as family	Follows and reuses the VTO terms
Oxudercin-ae (subfamily)	Not listed or classified	Not listed or classified	Classified as a subclass of Gobiidae (family)	Follows and reuses the VTO classification up to the lowest existing taxonomic terms covered (Family Gobiidae). Adopts NCBITaxon terms for Subfamily Oxudercinae onwards

3.2.2 Terms Validation

There are certain criteria for ensuring that the logical representation of the ontology terms are relaying proper meaning and definition, which can be captured by the semantic inference engine. The fish ontology in this study is validated for content, information flow, process flow, consistency, and assertion validation using two methods. To validate the ontology there are two methods used. The first method is automated where the whole process was done using Protégé inference engine such as FaCT++ (Tsarkov & Horrocks, 2006), Hermit (Glimm et al., 2014), and Pellet (Sirin et al., 2007). The second method was manual validation by human experts on fish and ontology development.

3.3 Ontology Evaluation

To evaluate the quality of the FO, we follow the Gruber method for ontology construction (Gruber, 1995). There are 5 criteria highlighted in this research which are clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment. Ontology clarity refers to how well the ontology model is defined, coherence refers to the ontology model consistency, and the extendibility refers to the ontology capability to be expanded and integrated. The ontological commitment can give a meaning of "a mapping between a language and something which can be called an ontology". Ontology modelers sometimes have a vague idea of the role each concept will play such as their semantic interconnections, within the ontology. If necessary, they can annotate new development ideas during the next update, which in turns increases its ontological commitment (Nicola et al., 2005). Encoding bias occurs when a representation choice is made for the convenience of notation or implementation. By minimizing encoding bias, knowledge-sharing agents may be implemented in different representation systems and styles of representation.

To measure the clarity level of the FO, the ontology definitions should be objective and independent of the social and computational context. To ensure the coherence quality of the FO, the definition of concepts given in the ontology should be consistent. While building the FO, the inferences drawn from the ontology must be consistent with its definitions and axioms. To further extend and simplify the coherence test for our ontology, we use the Ontology Debugger Tools from Protégé.

For extendibility evaluation, we evaluate the design of the FO pertaining to concepts and classification hierarchy represented as classes. The need for easy ontology extension is an important feature for the FO. It would be necessary to regularly update the existing ontology as new knowledge emerges regularly. For the low ontological commitment, we evaluate whether the ontology makes as few claims as possible about the domain while still supporting the intended knowledge sharing. For evaluating the encoding bias, we evaluate whether the ontology is independent of the issues of implementing language. Also, we check whether the conceptualization of the ontology is specified at the knowledge level and is independent of symbol-level encoding.

To strengthen the results of the FO evaluation, we use an online ontology evaluation tool named OOPS! Ontology Pitfall Scanner (OOPS) (Poveda-Villalón et al., 2014). OOPS uses a checklist to ensure that best practices are followed and that bad practices are avoided. The inventor created a catalog of bad practices and automated the detection of as many of them as possible (41 currently).

CHAPTER 4: RESULTS

The results for this research are broken down into several parts. There are 3 main parts in this study and each part is covered in subchapters below.

4.1 Fish Ontology

The results of creating the FO are further discussed in the following sections.

4.1.1 Fish Ontology Framework

The Fish Ontology (FO) consists of 652 classes (terms), and 27 object properties (relationships). There are 10 main classes which act as the core classes covering fish related and non-related terms within the FO structure. FO provides terms related to fish and infer species related information based on data that are fed to it. Current version of the FO is able to classify jawless fish, early jawed fish and living fossil fish. The FO contains 253 classes dedicated to fish studies and 38 classes related to fish sampling processes. Figure 4.1 shows the structure of some of the main classes in the FO and its lower level classes, while Table 4.1 give the statistic of imported classes and relationship in the FO.

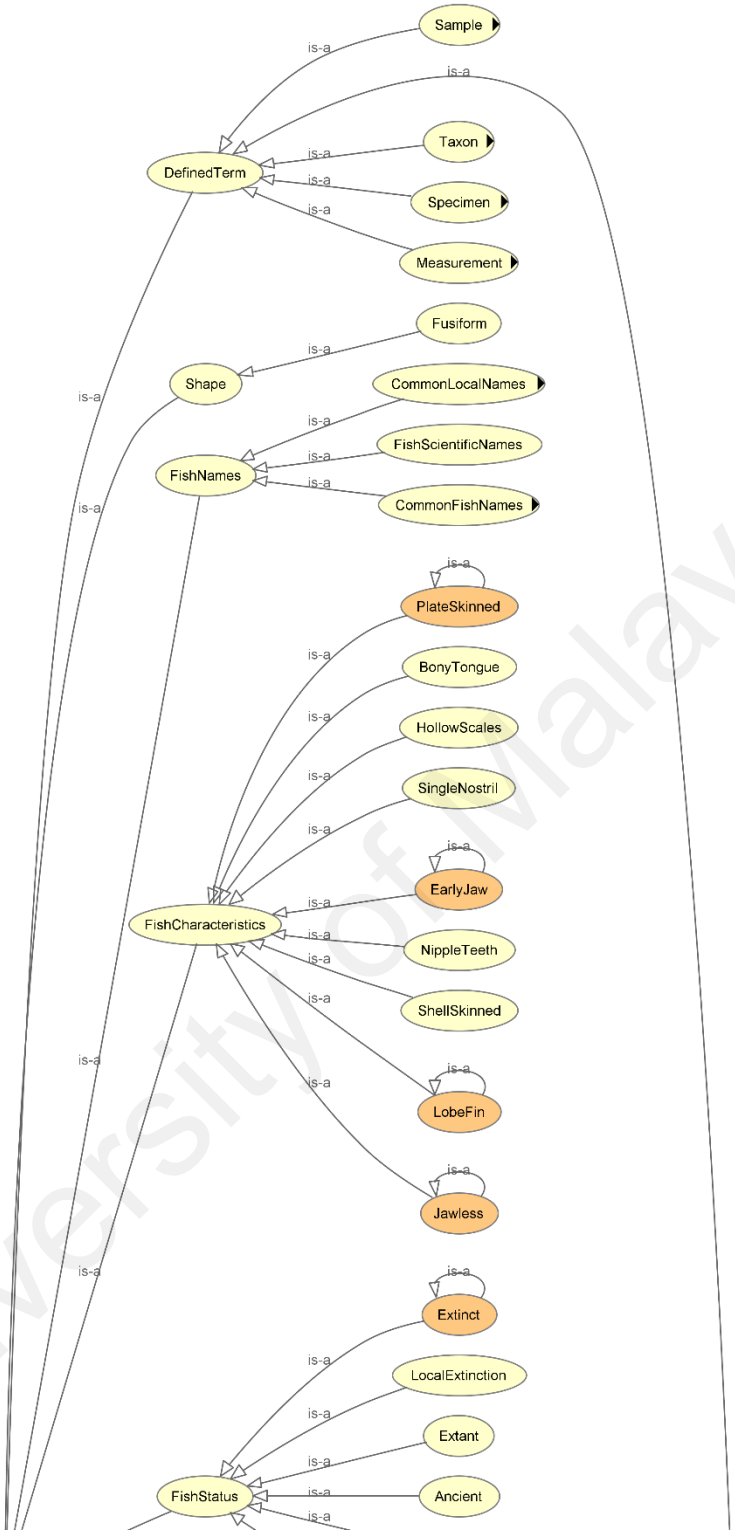


Figure 4.1: Structure of main classes and the subclasses of Fish Ontology. Yellow colored are normal classes while the orange colored are the classes with inferred properties.

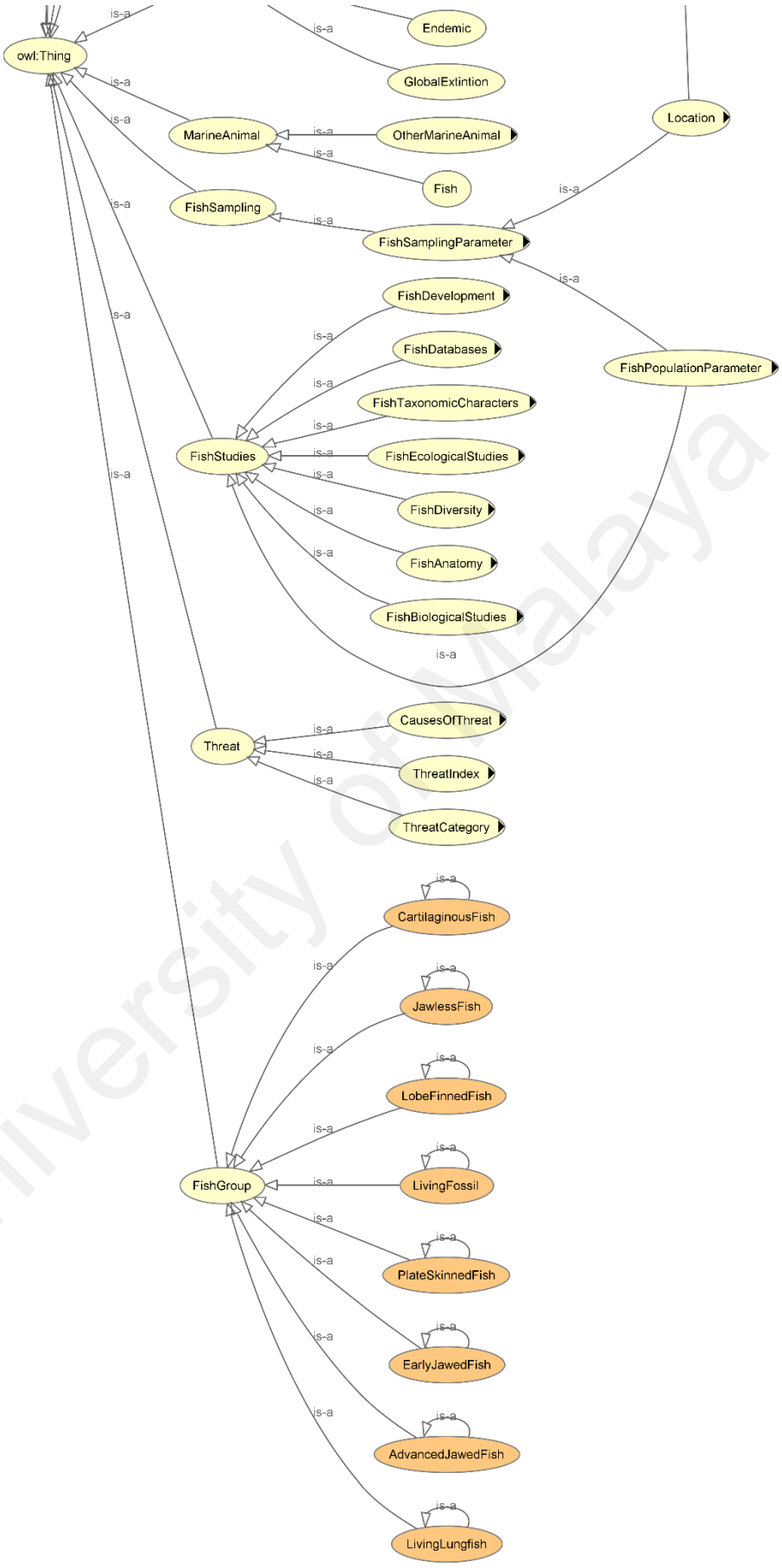


Figure 4.1: continued.

Table 4.1: Statistic of imported or integrated classes and properties.

Ontology or Standard	Number of classes
Zebrafish Anatomy and Stage Ontology (ZFA, ZFS)	2
Darwin Core	2
Vertebrate Taxonomy Ontology (VTO)	1345
NCBI organismal classification (NCBITaxon)	13
Total	1362

The FO reused 1345 VTO classes which are organized properly as the FO structure hierarchy model. For the "Taxon" class, it is organized in single inheritance, up to species level whenever possible, to increase the reasoning capabilities and expand its scope by further including relationship and annotations to the terms. This includes imported classes, which are linked to their respective class types. Each FO branch is organized hierarchically by means of the "is_a" (or subclass of) relationship, by appropriately placing it under a single root term. One relevant aspect of these classes is that they already have their own annotations in order to help understand the purpose. The FO framework have been uploaded to GitHub and can be accessed at the URL *https://raw.githubusercontent.com/mohdnajib1985/FishOntology/master/FishOntology.owl* or *http://www.essepuntato.it/lode/owlapi/reasoner/https://raw.githubusercontent.com/mohdnajib1985/FishOntology/master/FishOntology.*owl.

4.1.2 Fish Ontology Integration

To ensure integration with other ontology, it is imperative to properly reused the same terms, and keep the classes structure as similar as possible to the original ontology. As such, while creating the FO, all the possible terms structure for possible ontology integration are kept in mind to ease ontology integration. Figure 4.2 shows structure comparison between the VTO and the FO main classes and its subclasses to explain how other ontologies terms are imported into the FO using Protégé. While importing the desired terms into the FO, we retain the original structure of the terms taken from the VTO so that it will not change its real meaning.

4.1.3 Linking Fish Ontology with other databases.

One way of linking ontologies and databases is through the use of annotations. By using the tag “hasDBXref”, it is possible to link the desired terms with known database set. Figure 4.3 shows how the annotation is done in the FO so that it can be linked to other database sources. From the example, the terms in the FO are being linked to the PaleoDB, a database for fossils information.

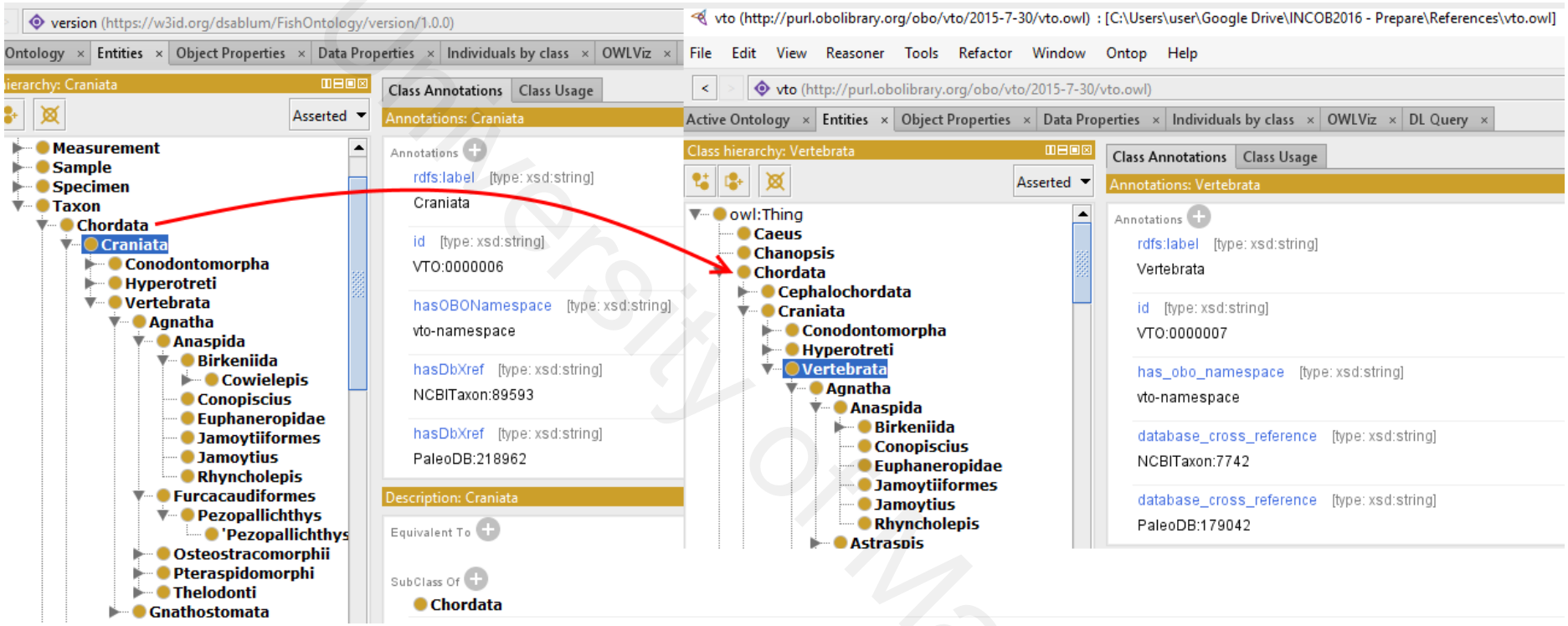


Figure 4.2: Structure comparison between the Vertebrate Taxonomy Ontology and the Fish Ontology main classes and its subclasses.

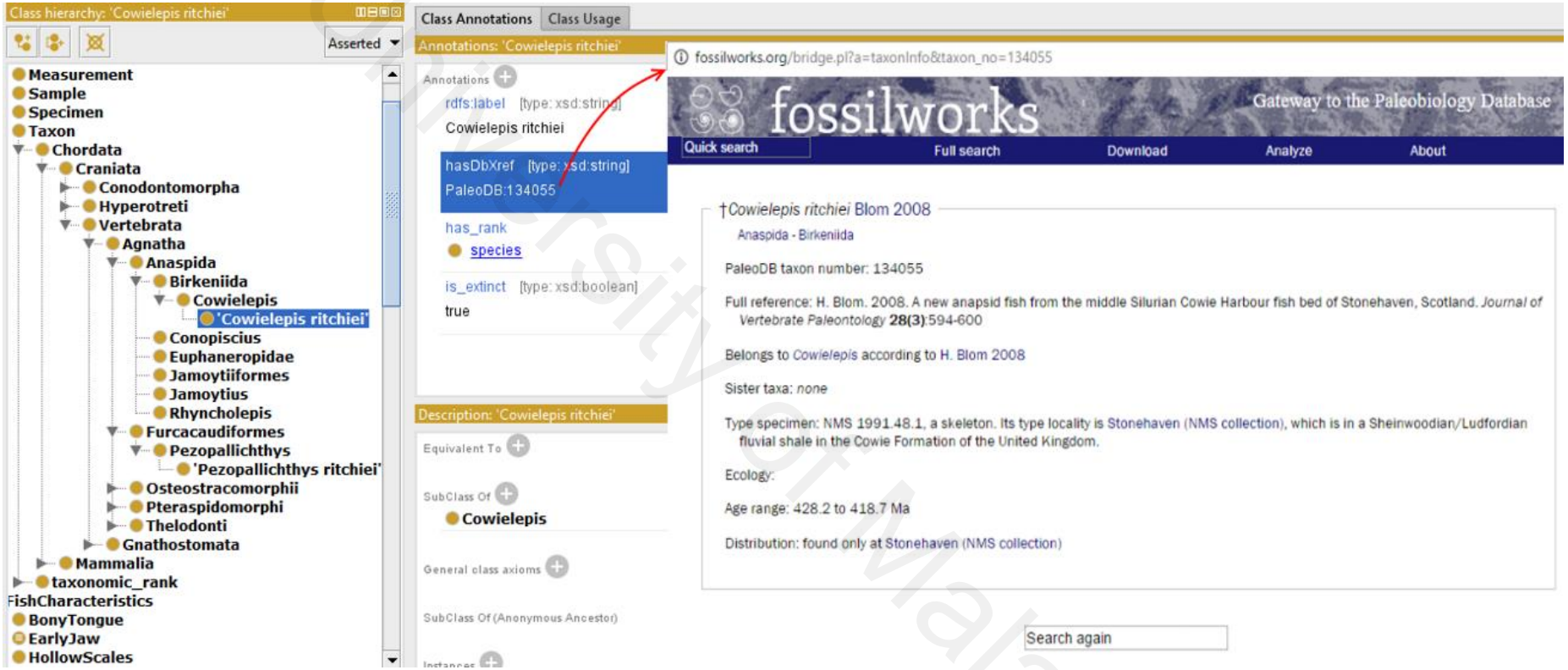


Figure 4.3: An example of linked annotation to map the Fish Ontology classes to the PaleoDB website.

4.1.4 Fish Ontology Relationships

As shown in figure 4.1 above, several classes have no direct relation to fish such as "defined_terms" and "threats". However, they are important nonetheless to further enhance the inferring capabilities of the FO. All of the classes in the FO have been observed for their usage, and only after careful consideration, are integrated into the ontology. The criteria for choosing the terms (discussed in the method section) ensures that the created FO is unique while capable of being integrated to other ontology. There are several ontologies or standard that have been adopted to the Fish Ontology (Table 4.2).

Table 4.2: Relationships in the Fish Ontology.

Property	Explanation	Examples
is_a	A subclass in OWL	Overharvesting is_a CausesOfThreat
hasRank (FO:0000097)	Describe a term which has a taxonomic rank	Carpet Shark hasRank of Orectolobiformes
isNameFor (FO:0000235)	Describe a name for some other class	FishNames isNameFor Fish
isGroupFor (FO:0000171)	Describe a group of some class	FishGroup isGroupFor Fish
isPartOf (FO:0000280)	Describe a situation where the class is part of something	PreflexionLarva isPartOf Larva

4.1.5 Inferencing Capabilities

The structure and the relationships discussed in the section above ultimately give the inferencing capabilities to the FO. As such, the FO can infer new information based on several restrictions that are fed to it. If there is a new specimen or sample that are added to the ontology while having the right parameter constraint, more information can be generated to determine the species of the fish. Figure 4.4 and 4.5 will further demonstrate the inference capabilities in the FO and show how inferred information is generated from a new sample or specimen based on metadata restriction.

4.1.6 Querying Capabilities

Fish Ontology supports several querying languages such as SPARQL, SPARQL-DL or SQWRL which are used primarily in querying RDF or OWL data mapping. Figure 4.6 shows several examples on how FO can be used to query data. As shown in the figure, not only does the FO allow querying its own content, it also provides a query result from inferred data from other ontology that is integrated into the FO, provided that proper querying tools are used. The query shown below is the results obtained after using the SPARQL-DL querying tools provided by Protégé.

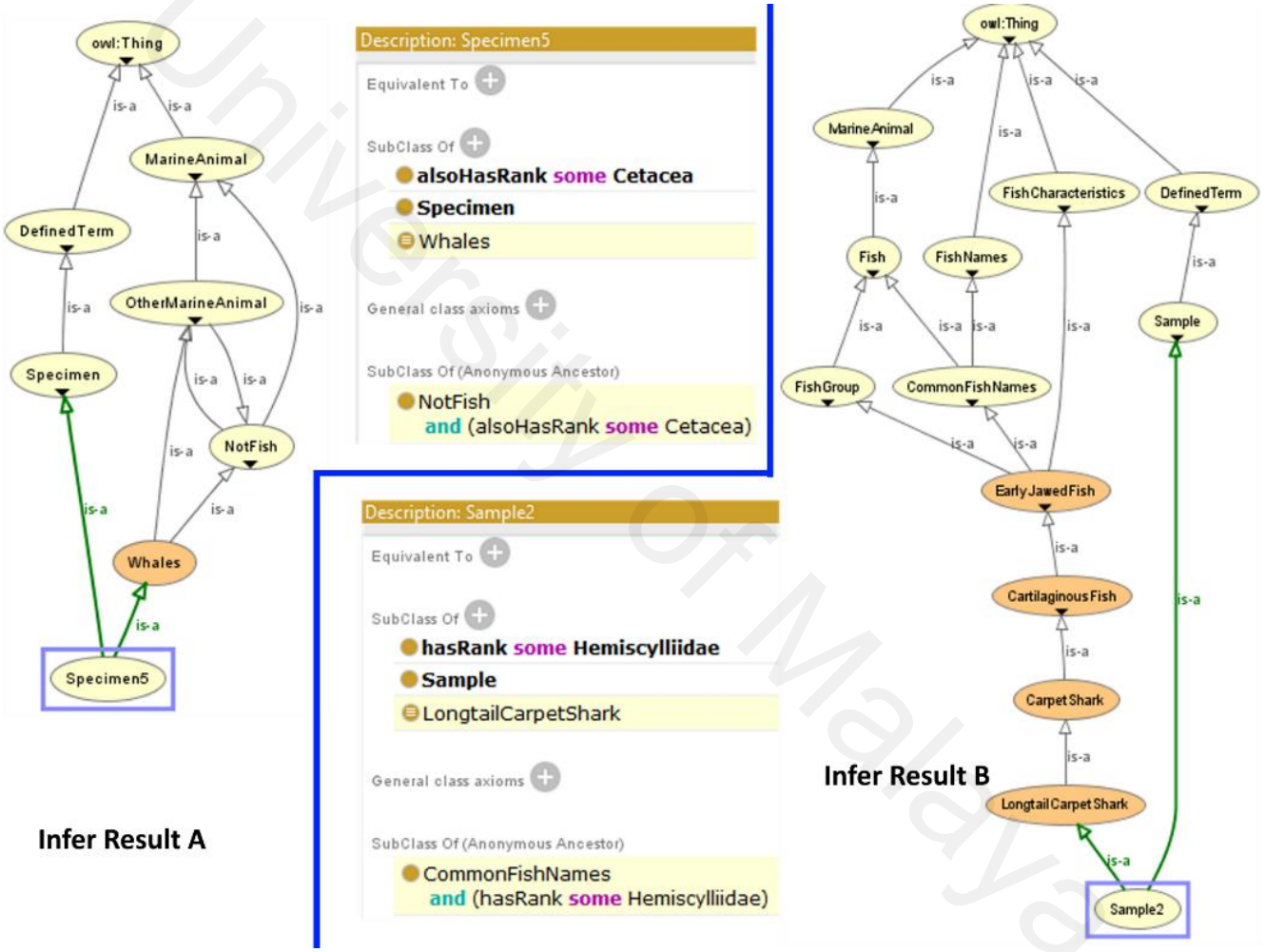


Figure 4.4: Inferencing capabilities shown through visualization of some classes in the Fish Ontology.

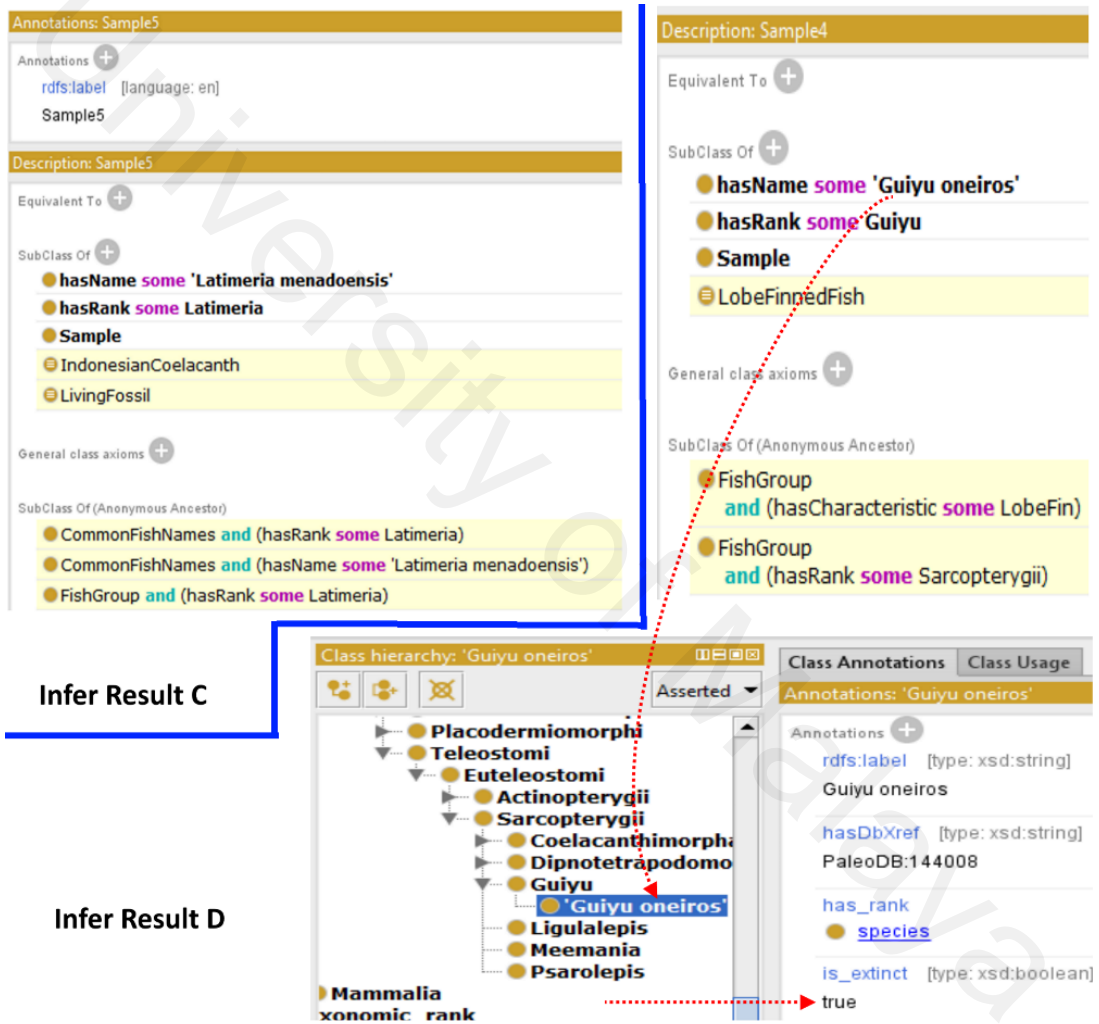


Figure 4.5: Results generated from the inference tools for some classes in the Fish Ontology.

Snap SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX fo: <http://mybiodiversityontologies.um.edu.my/FO.owl#>

SELECT * WHERE {
fo:Sample1  rdfs:subClassOf ?sub.
}
```

?sub
fo:Sample1

Query A

Snap SPARQL Query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX fo: <http://mybiodiversityontologies.um.edu.my/FO.owl#>

SELECT * WHERE {
fo:Sample1  rdfs:subClassOf ?sub.
}
```

?sub
fo:CartilaginousFish
fo:Sample
owl:Thing
fo:FishNames
fo:DefinedTerm
fo:EarlyJawedFish
fo:Fish
fo:FishGroup
fo:CommonFishNames
fo:Sample1

Query B

Figure 4.6: Results generated from querying some statement in the Fish Ontology. Query A shows the results of querying the class "Sample1", retrieving all of its subclasses, without using any inferences. Query B shows the same query with different results while using inference tool in Protégé.

4.2 Fish Ontology Evaluation

There are 5 parts for the evaluation section, explained in the subchapters below.

4.2.1 Clarity

In the FO, all the definitions are stated in such a way that the number of possible interpretations of a concept would be restricted. The clarity test results for the FO are divided into 6 parts which are:

1. No Cardinality Restriction on Transitive Properties
2. No Meta-Class
3. No Subclasses of RDF Classes
4. No Super or Sub-Properties of Annotation Properties
5. Transitive Properties cannot be Functional

Results for tests 1 and 4 are shown in Figure 4.7 below. Since fish data are large in volume, there is a need to add more data over the time. As such, there is no cardinality restriction assigned to any transitive properties in the FO. Figure 4.7 also shows that the transitive properties are also not functional because it relates to more than one instance via the property. As for tests 2, 3 and 4, Figures 4.7 and 4.9 show that there are no meta-classes, no properties with a class as a range, and no sub-classes of RDF classes in the FO. Furthermore, since we used the Protégé as the development tool, all the 5 tests are automatically filtered, because these criteria are automatically flagged in the latest Protégé version.

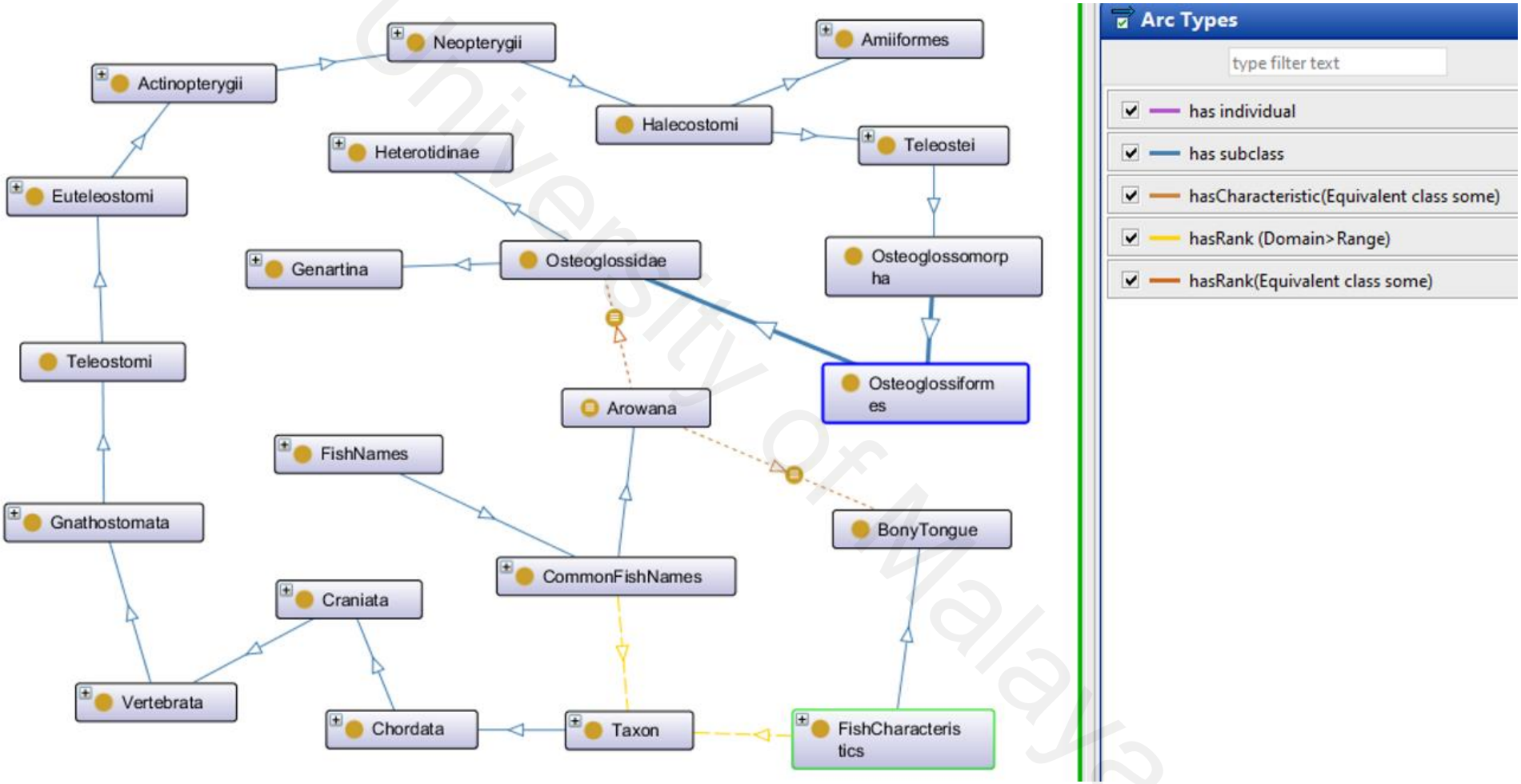


Figure 4.7: Results for clarity tests (1, 2, 3 and 4).

4.2.2 Coherence

The first main result of the coherence test can be seen in Figures 4.4, and Figure 4.5. Here, we can see that most of the inferred terms from the ontology are consistent with their definition and axioms. As an example, in Figure 4.4, when the FO inferred that Specimen5 is a whale, it also inferred that it is not a fish, and has the correct taxon rank. The formal part of the FO is checked by following these 6 consistency criteria listed below and ensuring that all return true:

1. Domain of a Property should not be empty
2. Domain of a Property should not contain redundant Classes
3. Range of a Property should not contain redundant Classes
4. Inverse of Symmetric Property must be Symmetric Property
5. Inverse Property must have matching Range and Domain

The usage of software (Protégé) forces the user to always be wary about an empty domain, redundant classes, and properties. As such, tests 1 to 3 are achieved and can be further viewed through the ontology itself via the link URL provided in the last paragraph of chapter 4.1. For test 4, we provide an example of the property isSimilarTo. The class CosmoidScales is related to the class PlacoidScales via the isSimilarTo property. Then we can infer that PlacoidScales must also be related to CosmoidScales via the isSimilarTo property. Figure 4.8 shows the results of coherence test using the Ontology Debugger Tool from Protégé. The coherence test from this tool checks for possible faulty axioms. The ontology passed the coherence test provided by this tool. Figure 4.9 shows the results for test 5 showing that the properties hasCharacteristic and isCharacteristicFor have matching range and domain.

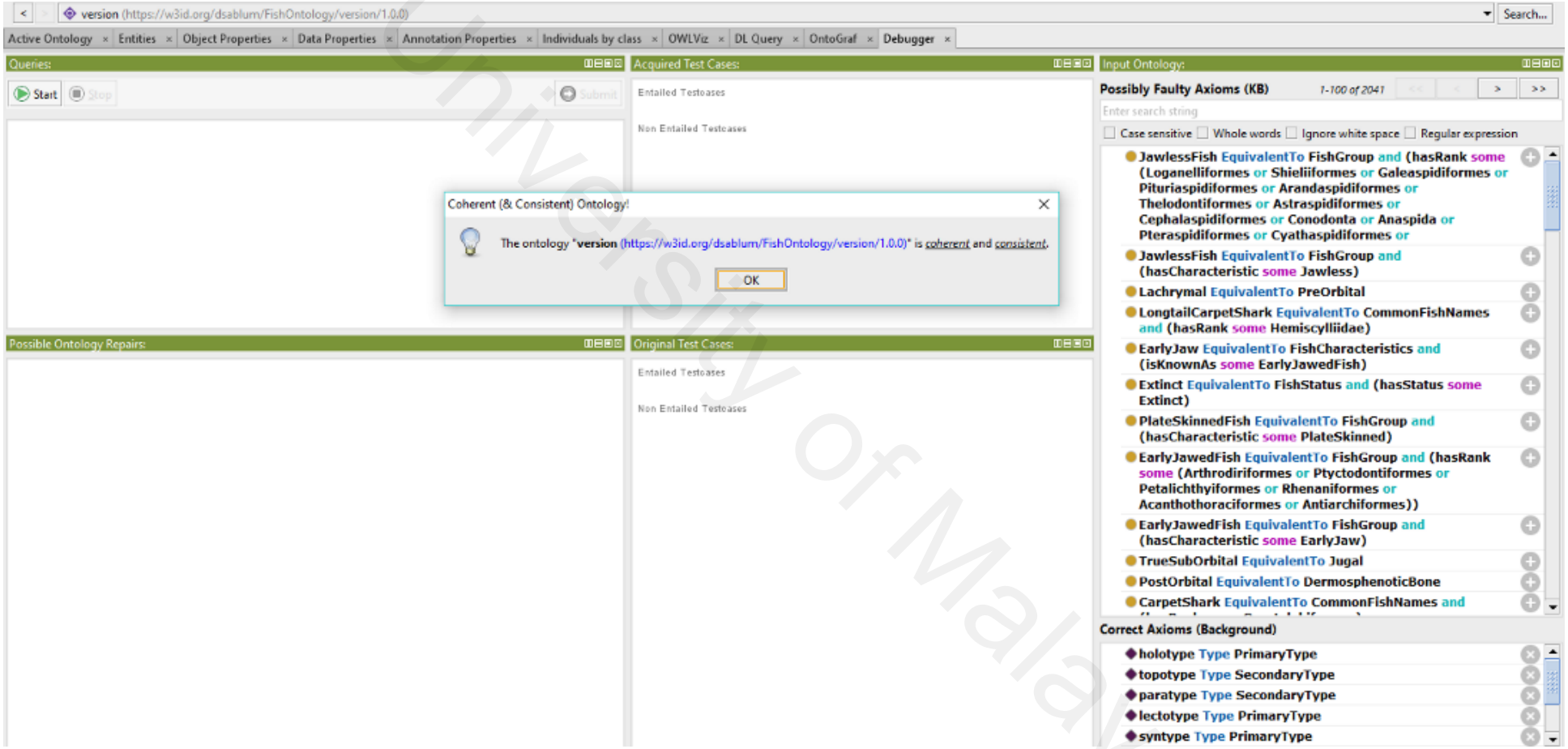


Figure 4.8: Results of the coherence test using Protégé Ontology Debugger tool.

4.2.3 Extendibility

Table 4.1, Figure 4.2, and Figure 4.3 show the extendibility of the FO. Since the first design, we have considered integrating terms from other ontologies into the FO. By placing any related concepts derived from other generic concepts in its class hierarchy, the FO represents information that defines a fish specimen, linking it with terms from other ontologies. Creation of classes and annotations that may be useful for future integration such as "genetic content" will further enhance FO's extendibility.

4.2.4 Low ontological commitment

Since the FO reuses existing concepts (from books, databases and other ontology) and proposes only a few new concepts, it has low ontological commitment. The low ontology commitment makes the FO more extensible and reusable. Also, since most of the new concepts are from notable books and published journal articles (Chong et al., 2010; Helfman et al., 2009; Last et al., 2010; Nelson, 2006), the concepts will be more widely accepted among the user community.

4.2.5 Minimum encoding bias

The choices of using OWL as the representation language and to stick with terms from books, database, and related ontology (shown in Table 3.2, Table 3.3, and Table 4.1), are intended to reduce the encoding bias. Furthermore, Figure 4.10 shows that there are no errors regarding encoding bias.

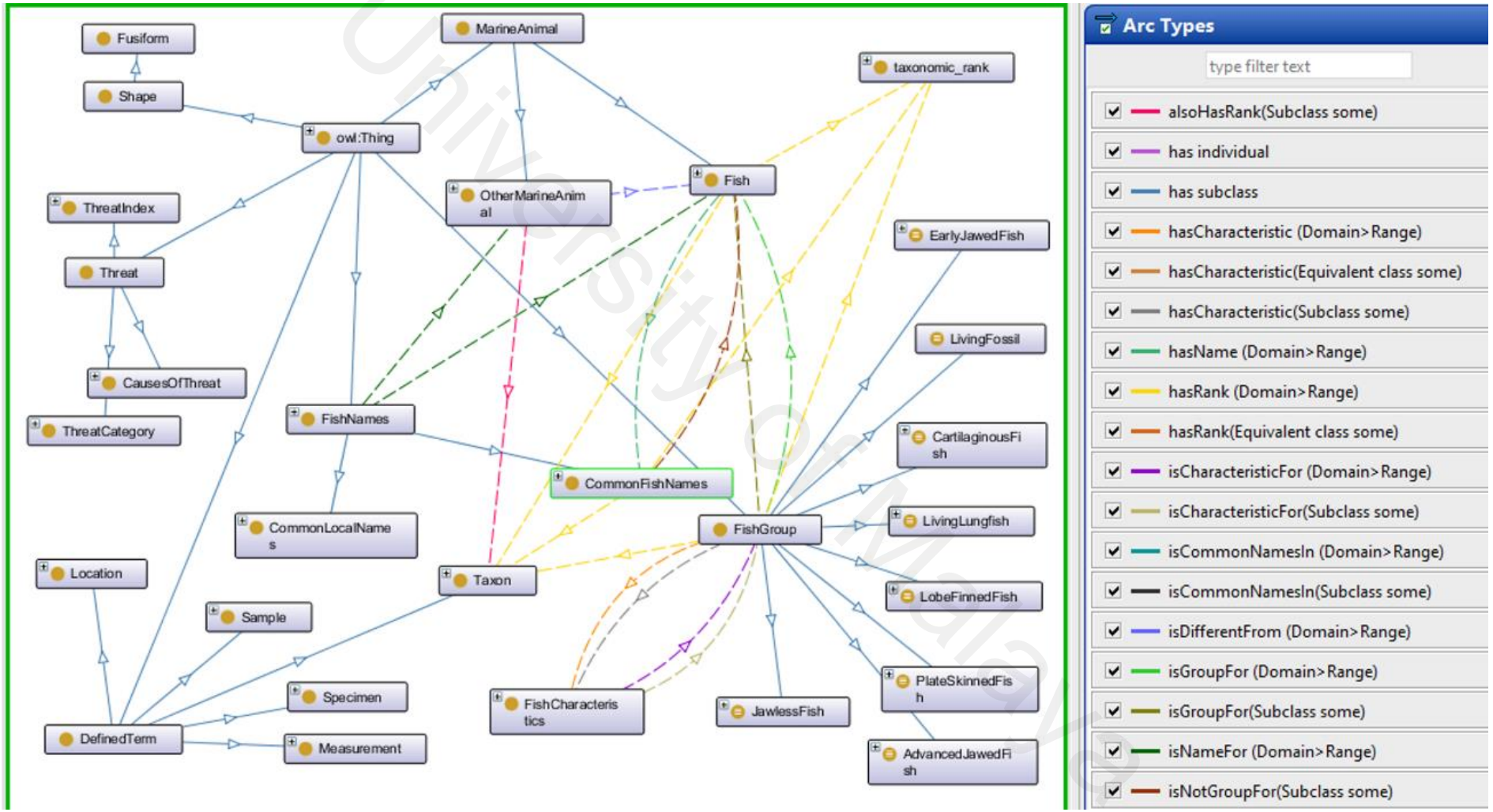
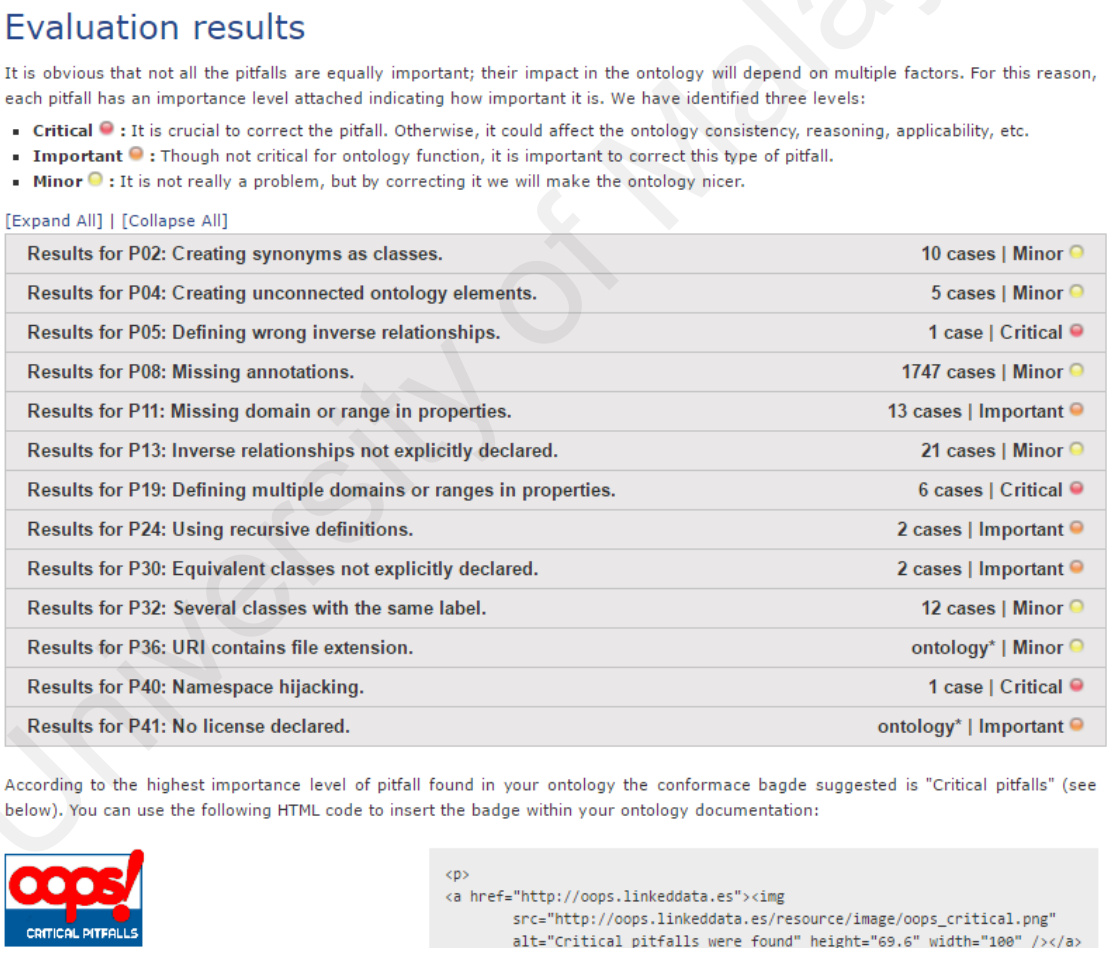


Figure 4.9: Results for clarity test (3, 4, and 5), coherence test (5).

Ontology Pitfall Scanner evaluation

The evaluation of the FO using the Ontology Pitfall Scanner (OOPS) tools is shown in Figure 4.10. There are 1794 cases listed in the minor pitfall categories, 19 cases in 4 important pitfall categories, and 11 cases in 4 critical pitfall categories. Compared to the ontology debugger tools in the Protégé, there are many error flags that can be found in the FO by using OOPS. However, most of them are minor, and the important and critical pitfalls problems are mostly caused by the same features in the FO, and will be further elaborated in discussion section.



Evaluation results

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

- **Critical** : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** : It is not really a problem, but by correcting it we will make the ontology nicer.

[Expand All] | [Collapse All]

Results for P02: Creating synonyms as classes.	10 cases | Minor
Results for P04: Creating unconnected ontology elements.	5 cases | Minor
Results for P05: Defining wrong inverse relationships.	1 case | Critical
Results for P08: Missing annotations.	1747 cases | Minor
Results for P11: Missing domain or range in properties.	13 cases | Important
Results for P13: Inverse relationships not explicitly declared.	21 cases | Minor
Results for P19: Defining multiple domains or ranges in properties.	6 cases | Critical
Results for P24: Using recursive definitions.	2 cases | Important
Results for P30: Equivalent classes not explicitly declared.	2 cases | Important
Results for P32: Several classes with the same label.	12 cases | Minor
Results for P36: URI contains file extension.	ontology* | Minor
Results for P40: Namespace hijacking.	1 case | Critical
Results for P41: No license declared.	ontology* | Important

According to the highest importance level of pitfall found in your ontology the conformace bagde suggested is "Critical pitfalls" (see below). You can use the following HTML code to insert the badge within your ontology documentation:


```
<p>
<a href="http://oops.linkeddata.es"><img
        src="http://oops.linkeddata.es/resource/image/oops_critical.png"
        alt="Critical pitfalls were found" height="69.6" width="100" /></a>
```

Figure 4.10: Results of evaluation using the Ontology Pitfall Scanner tool (Poveda-Villalón et al., 2014).

4.3 Fish Ontology Portal

The Fish Ontology Portal is a prototype web portal created for the purpose of accessing all the information that are stored in the Fish Ontology. All the data in the portal are queried using SPARQL. The purpose of creating this prototype is to test the ontology capabilities in a real world application. Its creation may help researchers, academicians, and students to monitor and view species occurrence, fish information, and fisheries activities in any area of interest. There are 2 prototypes created in this research; one is created using Apache Jena, and the other using SESAME. Here we show the results of both frameworks with regards to their capabilities, strength, and weakness.

The first prototype portal (Figure 4.11) in this study was developed using the Apache Jena framework as its system environment. Apache Jena provides all the necessary tools to retrieve the data within the ontology and to add more data or new terms using by using the built-in querying capabilities. Feasibility and performance test were carried out on this prototype. In figure 4.11, number 1 shows the Fish Ontology Portal main page, number 2 shows the Search results function demonstration for alphabet “A”, number 3 shows the specimen list for a species, number 4 shows the view page for a species, number 5 shows the morphological details view of a specimen, number 6 shows the catch details of a specimen, number 7 shows all the citation details for a species, number 8 shows the specimen editing main page, number 9 shows the catch details editing page, number 10 shows the other minor details editing for a specimen, number 11 shows the researchers editing main page, and finally number 12 shows the image gallery. There are several functions of the portal summarized below:

1. Reading all the data from the Fish Ontology OWL files.
2. Inserting new fish data into the Fish Ontology OWL files.

3. Searching for fish information using several parameters such as name, location, etc.
4. Prototype semantic search function which can find new information about a species in other ontology web API.
5. Image gallery of species within the owl files.

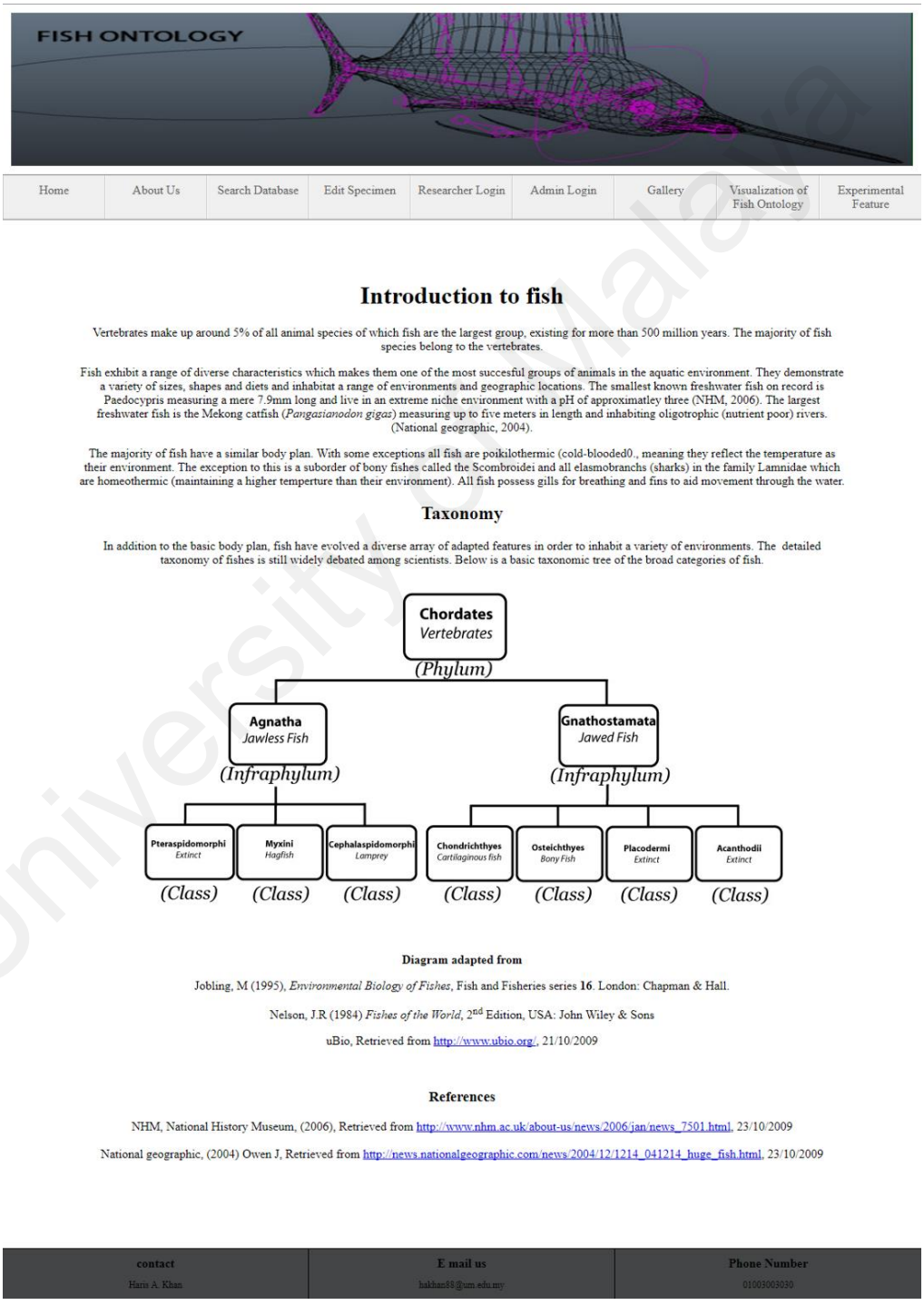


Figure 4.11: Front page of Fish Ontology Portal.

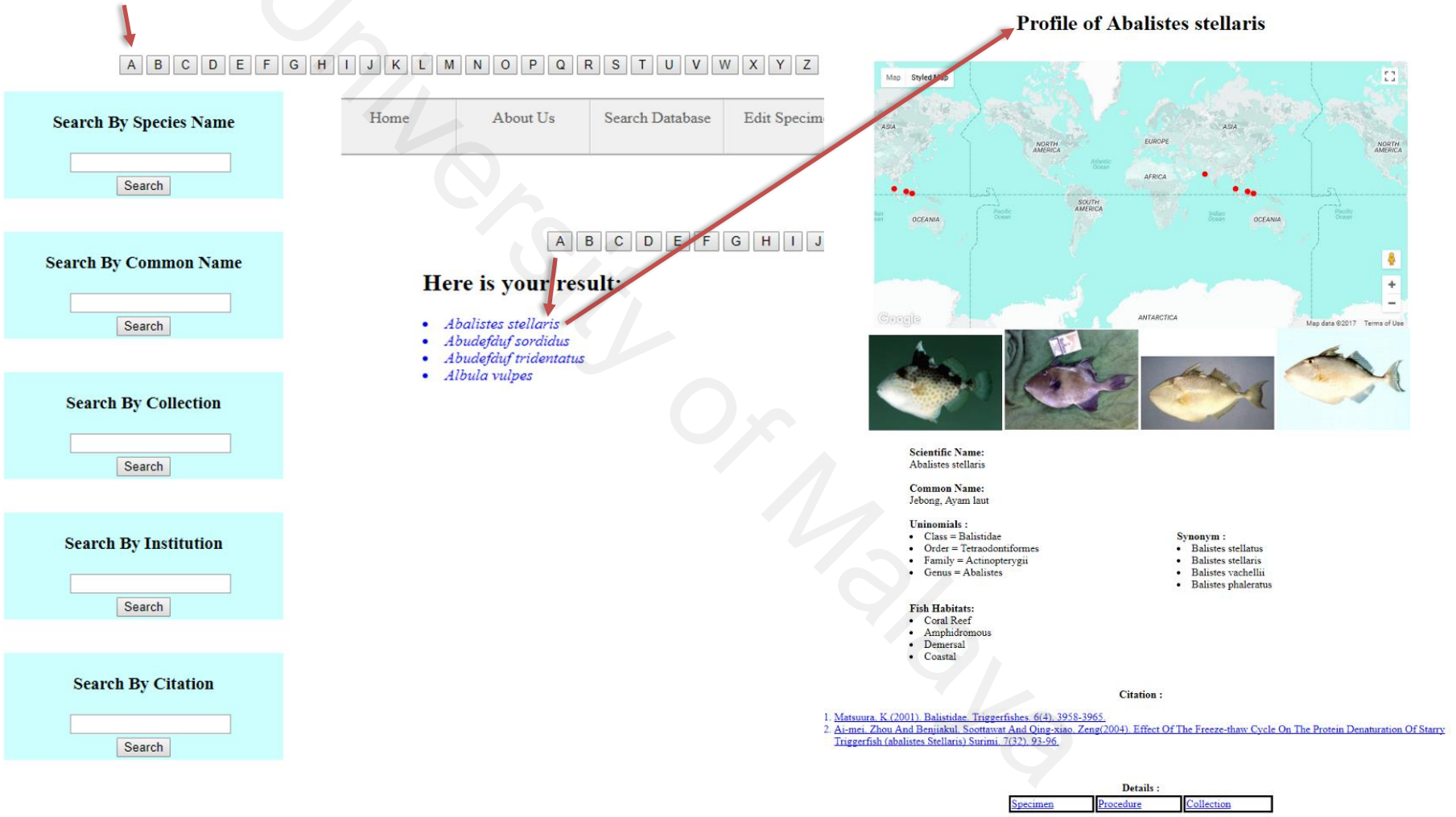


Figure 4.12: Search function of Fish Ontology Portal.

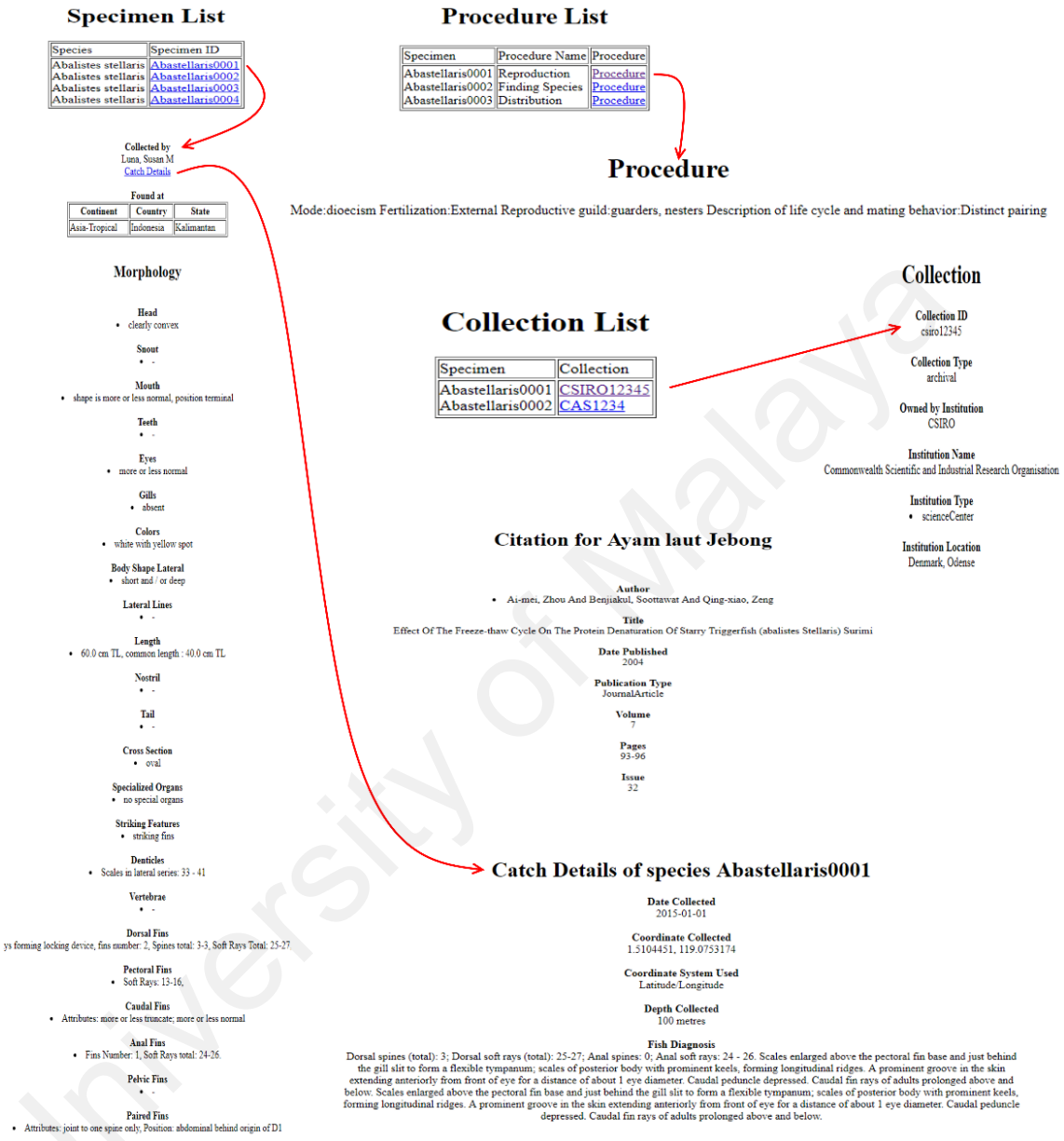


Figure 4.13: Fish and specimen details.

Figure 4.14: Updating specimen in Fish Ontology Portal.

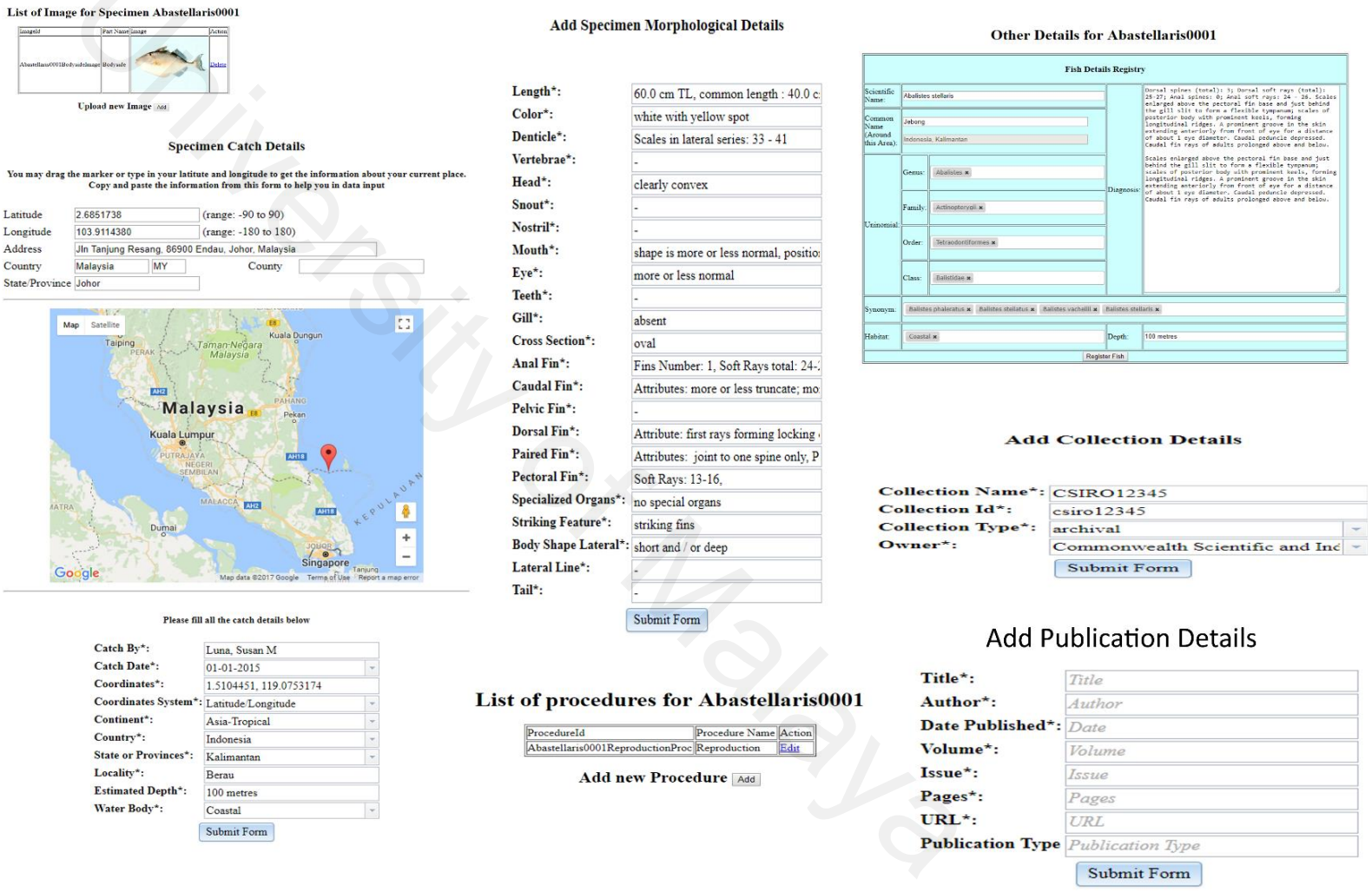


Figure 4.15: More details on specimen update in Fish Ontology Portal.

CHAPTER 5: DISCUSSION AND CONCLUSION

In this chapter, several important points in this study are discussed thoroughly such as current issues related to the study, strength and weaknesses of the Fish Ontology, and finally, its future directions and enhancement.

5.1 Ontology and portal creation

In this study, a Fish Ontology is proposed. This ontology is a general-purpose ontology that allows integration of domain-specific biodiversity ontologies containing standard terms and relationships. The design of the FO is flexible enough to accommodate any biodiversity ontology containing data or knowledge about fish. Even in cases where integration can be difficult, the FO can be tweaked in order to incorporate new biodiversity-related ontology. One example is linking the FO to the MarineTLO (Tzitzikas et al., 2016), which is an upper-level ontology for marine species. The MarineTLO does not have a class named "Fish" that can map to data from the FO. However, since the MarineTLO provides classes of taxonomic rank such as "Species" and "Genus", and related classes such as "MarineAnimal" and "Specimen", the FO can then create the necessary annotations to link these classes. The same can be done to ZFIN (Sprague et al., 2003; Van Slyke et al., 2014) which contains "zebrafish anatomical entity" and "Stages" as main classes; the FO can generate main classes such as "FishAnatomicalEntity" and "OtherStagesTerminology".

There are other resources that model animal taxonomy which can be used to build the FO, such as the NCBITaxon (Federhen, 2011) which is an automatic translation of the NCBI taxonomy database into .obo or .owl format (Federhen, 2016), However, the NCBITaxon differs from the FO where it models only the taxonomic ranks without fish characters and nomenclature. The NCBITaxon also has a different hierarchical organization and definitions compared to the VTO which is used as the main reference

for taxonomic characters and rank in the FO. The VTO is directly imported to the FO because it is built following several taxonomic resources, including the NCBI Taxonomy, the Paleobiology Database (PaleoDB) (Alroy et al., 2012), and the Teleost Taxonomy Ontology (TTO) (Dahdul et al., 2010; Midford et al., 2010), which suits the need of the FO for a comprehensive fish taxonomy information. One of the most distinctive values of the VTO compared to others is its broad taxonomic coverage of the vertebrates. The NCBITaxon however, excludes many extant and nearly all extinct taxa, while largely include only species associated with archived genetic data, complemented by data from the PaleoDB and the TTO to provide an authoritative hierarchy and a richer set of names for specific taxonomic groups (Midford et al., 2013). Having said that, we incorporate taxon ranks which are covered by the NCBITaxon but not the VTO, such as "*Protanguilla palau*" and "Oxudercinae". In general, we follow the information such as synonym, name, fish grouping, and group rank, and fish, fisheries and fish studies related terms provided in the book (Helfman et al., 2009) as the main structure of the FO and adopt the usage of the VTO for taxonomic hierarchy, taxonomically related information, and terms related to taxonomic rank.

In most cases, the taxonomy of the VTO is followed as it is a regularly updated ontology. One exception is the class "Mammalia" which the VTO classified as under "Sarcopterygii" (meaning that it is derived from fish). There are differing views on this specific classification and we opted not to follow this specific structure provided by the VTO. The use of adopted terms and concepts from our main references (Helfman et al., 2009) is further clarified with domain experts (Amy Y. Then, Chong V. Ching) in order to represent and map the appropriate contents to reflect the diverse aspects of fish. The new terms are checked for its suitability to be adopted as a standard vocabulary for fish scientists. Proposing new vocabulary in biodiversity is not uncommon since ontologies in this domain are presently insufficient and many are under development. Available

standard vocabulary is not comprehensive enough to cover all the terms needed to make an ontology in the fish domain. In most cases, new terms must be proposed based on the rationale utilized in the ontology. One such example is that of Hymenoptera Anatomy Ontology (Yoder et al., 2010), where new terms had to be proposed to expand the ontology (Seltmann et al., 2012, 2013).

Fish represents the most diverse vertebrate group on Earth; hence coverage of all possible terms and parameters for the fish domain by a single ontology is not possible. The current FO version covers the terms for fish domain which are not well described by other ontologies, particularly those related to automatic classifications, annotations, and relations. There are however other parameters rarely used outside this domain, such as "FishDatabases" which shows known databases for fish, or "GasBladder" which is a specific organ for "Actinopterygii". Thus, there is a need to develop ontologies that cover these specific fish concepts and parameters while reusing relevant terms from existing ontologies in related domains.

Regarding ontology evaluation, there are reasons a number of errors were flagged by the Ontology Pitfall Scanner tool (OOPS) but none can be detected by using the tools from Protégé. The most apparent reason is because the scope of evaluation for both methods are different. In Protégé, only the classes and its relationship structures created in the ontology are being evaluated, while in OOPS, the classes, relationships, mapping and future integration problems are being evaluated, giving different results. One of the most important features in the FO is reusing of terms from other ontologies to reduce term redundancy in global usage. As such, many terms and structures related to fish and fisheries are taken from other ontology such as the VTO, with proper indications and reference that they are taken from its source. The idea is to reduce terms redundancy in global usage. However, since most of the terms are directly used in the FO, the OOPS

tool flag these occurrences as critical errors such as “P24: Using recursive definitions”, “P32: Several classes with same labels”, and “P40: Namespace hijacking”.

Other pitfalls such as P02, P04, P08, P11, P13, P30, P36, and P41 (refer to Figure 4.10) are considered acceptable since there are constantly new items to be added to the ontology along with the necessary annotations, relations and property constraints. As for the pitfall “P19: Defining multiple domains or ranges in properties”, this is usually due to how the ontology is modelled. Unlike a typical ontology that use inferring capabilities to discover new relationships, we also use the inferring capabilities for automated fish species recognition. Therefore instead of using 1 to 1 relationships for the domain and range to restrict the use of the property, the usage of the property is enlarged so that it is more reliable for automated species discovery.

There are also some issues encountered during the Fish Ontology portal creation. Issues occur when the dataset provided by the fish expert has different names, although they have the same meaning. Various terms have been used for naming fields with same meaning thus it needs to be rechecked so that the field name, their abbreviations, and their short terms are matched with the data sources to ensure standardization. The need for standardization was previously neglected, often not fully implemented, and are not thoroughly pushed, especially around the 1980s. Furthermore, data collection around that time is based on researchers’ own research requirements and there is no further interest to share the raw data with other researchers or the scientific community. Data added to the FO need to be ensured so that it suits the needs of the scientific community and useful for research and evaluation. Correct and accurate data is important to the fish and fisheries community to further expand the information network and help the community to grow.

As for the development tools, apart from Apache Jena, we have tried other ontology development framework, such as SESAME (now known as RDF4J). However, we faced some difficulties in adding data using the framework due to a couple of reasons. Unlike Apache Jena which is heavy-weighted, the RDF4J is quicker at extracting data and more light weighted. While testing this framework, we also noticed that Sesame is able to query data significantly faster. This is because of the simplicity of its framework where RDF4J support two query languages (SPARQL and SeRQL) compared to Apache Jena with 3 query languages (SPARQL, SWRL, and SQWRL). It has many other functions that are not supported by Apache Jena such as adding indexing and query capabilities to all compatible stores. However, SESAME did not provide Full OWL editing capabilities, which makes us choose Apache Jena since it covered most of the needed functions. As far as we are aware of, the Apache Jena framework is more robust, and has many capabilities which are not available in SESAME framework. On the contrary, SESAME framework provides more speed and simplicity in terms of search function and ease of use. Table 5.1 contains the advantages and disadvantages of portals which were created using Apache Jena and Sesame frameworks.

Table 5.1: Difference between Apache Jena Framework and Sesame Framework.

Functionality	Framework	
	Apache Jena	**SESAME**
Load and insert data	Able to fully load and insert data from .owl file without conversion	Can fully load and insert data from .owl file but need to convert the owl file to a flat file database.
Framework environment purposes	Apache Jena is a Java framework for building Semantic Web and Linked Data applications.	RDF4J is a Java framework for processing RDF data, supporting both memory-based and a disk-based storage.
Accessibility	Can be accessed using Fuseki, Jena RDF API, RIO, and SPARQL	Can be accessed by Java API, RIO, Sail API, SeRQL, Sesame REST HTTP Protocol, and SPARQL
Language Support	Can support only Java Programming language	Can support Java, PHP, and Python programming languages.
Querying Speed	Speed wise, Apache Jena takes a bit of time querying inferred information, and it support the usage of SPARQL-DL.	Speed is way faster than Apache Jena however only allow SPARQL querying and does not support Description Logic inferring and cannot support OWL2.

There are other ways to implement semantic web technologies to a database set such as google knowledge graph which can handle linking information in the web as easy as just mapping each terms of other ontology or URL to the terms of interest in your portal. However this knowledge graph does not support querying using SPARQL query language, which is the main feature of OWL file format.

Ontology tailoring is computationally expensive, partly because of the size of the ontologies and also partly because of the complexity of the requirements of the user. Deriving Tailored Ontologies from large base ontologies enables individuals to use only specific parts of the ontology for their daily use. Most user applications only require particular aspects of the ontology as they do not benefit from the overabundance of semantic information that may be present in the ontology. Ontologies may be small, containing just a few concepts and relationships or they may be ever expanding, containing many millions of concepts and relationships. Ontologies are becoming popular largely due to what they promise: a shared and common understanding of a domain that can be communicated easily between people and applications.

5.2 Current Strength and Weakness

Data representation in the form of an ontology allows the linking of information by using semantic web applications. As shown in the results, the FO currently is the first biodiversity-related ontology capable of providing automated taxon information based on specimen or sample metadata constraint. It can provide fish information and description to fish-related terms such as extinction status, databases, taxonomic rank, and names (scientific, common, local). The current version of FO can classify jawless fish, early jawed fish and living fossil fish. Furthermore, it has the link to several published databases such as FishBase and PaleoDB which enhances the information for the terms in the FO. Moreover, it can also be used to prepare captured and observed fish specimen data, mapped and structured in a way that the meaning is expressed in a machine-understandable format.

Additionally, the current version of the FO can utilize specimen grouping and characteristics to determine whether the specimen is a fish or otherwise, provide taxonomic information and heredity of a characteristic rank, determine conservation

status, evolutionary status (ancient or modern) and type (ancient species is a jawless fish). This version uses simple character classification where the user provides the necessary character for the specimen. As an example, the user can specify that "Sample 1 has the characteristic of Plate Skinned", and manually add the characteristic of "Plate Skinned" into the FO. We believe the ideal version should contain anatomical and phenotype data from several classes in ontology such as "Anatomical Characteristics", "Meristic Characteristics", "Molecular Characteristics", and "Morphometric Characteristics" and these features will be included in the future. These classes can be useful for pattern recognition, and species taxon recognition studies. The power of the FO lies in its ability to automate group classification, and ability to link the terms used by fish domain researchers, and other researchers outside the domain.

The weakness of the FO lies in its position as a newly published ontology. Hence, the usage of the ontology is low and there might be little responses on how well its performance in tackling fish-related issues. Furthermore, the number of databases that it is linked to is still limited and there is still room for it to be linked with other ontology to increase its granularity. The current version of the FO also is yet to cover all parts of fish-related terms such as fish aging process, or fish sampling process properly.

5.3 Evolution and Future Directions

The FO have been through several drastic changes in the structure before it was finalized into its current version. The first version of FO is created purely based on TDWG LSID terms and only model the structure of fish taxonomy and its anatomical entity. The first version considered all the necessary terms integration, but no proper linking were made to the ontology in order for it to fully emulate the semantic web experience. Figure 5.1 to Figure 5.4 show the images of the all of the previous versions of the FO which has undergone many amendments over the period of the study. In the

first version, V1, shown by Figure 5.1, the most used terms in fish and fisheries area are incorporated in the design such as Specimen, TaxonRank, TaxonName, OccurrenceRecord, Morphology, Collection and DigitalImage. In the second version, V2, shown by Figure 5.2, further improvement was made in the TaxonName area and several terms of different ontology were incorporated such as VSAO:anatomical structure and scale. In the third version, V3, shown by Figure 5.3, the morphological part was expanded and several adjustments were made to the relationship between the classes. In the fourth version, V4, shown by Figure 5.4 more terms were added to expand the morphological features.

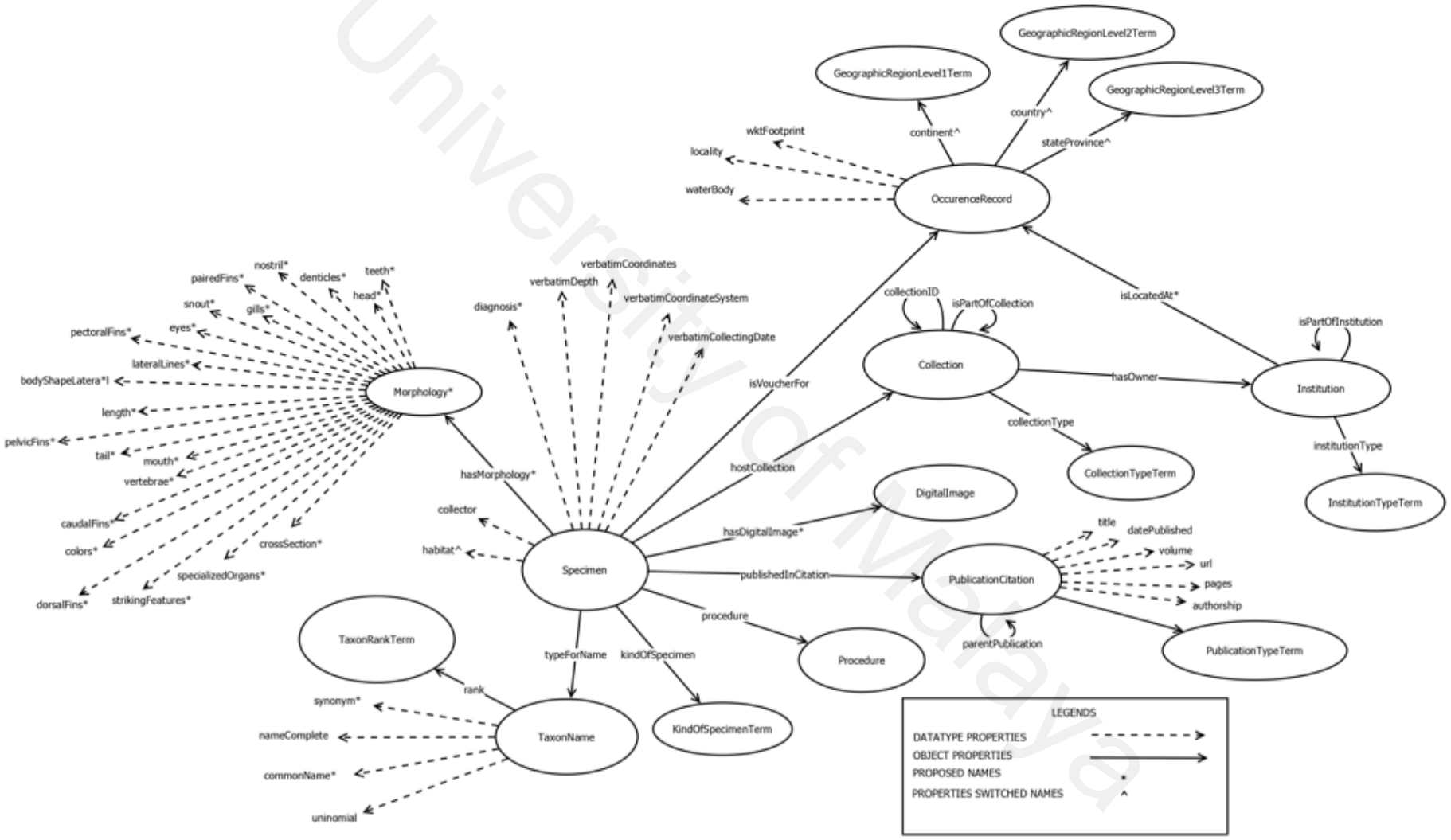


Figure 5.1: First version of Fish Ontology (V1).

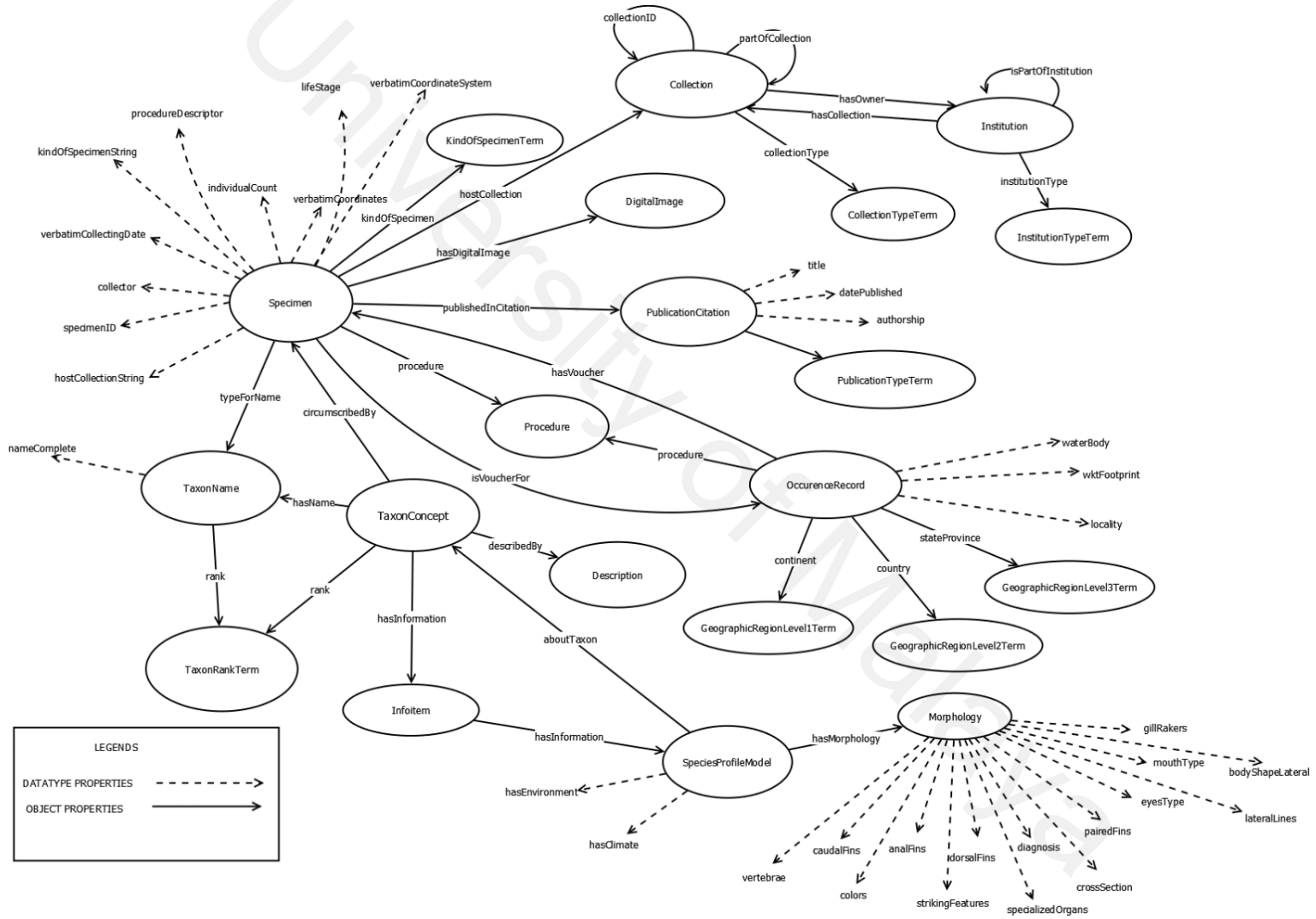


Figure 5.2: Second version of Fish Ontology (V2).

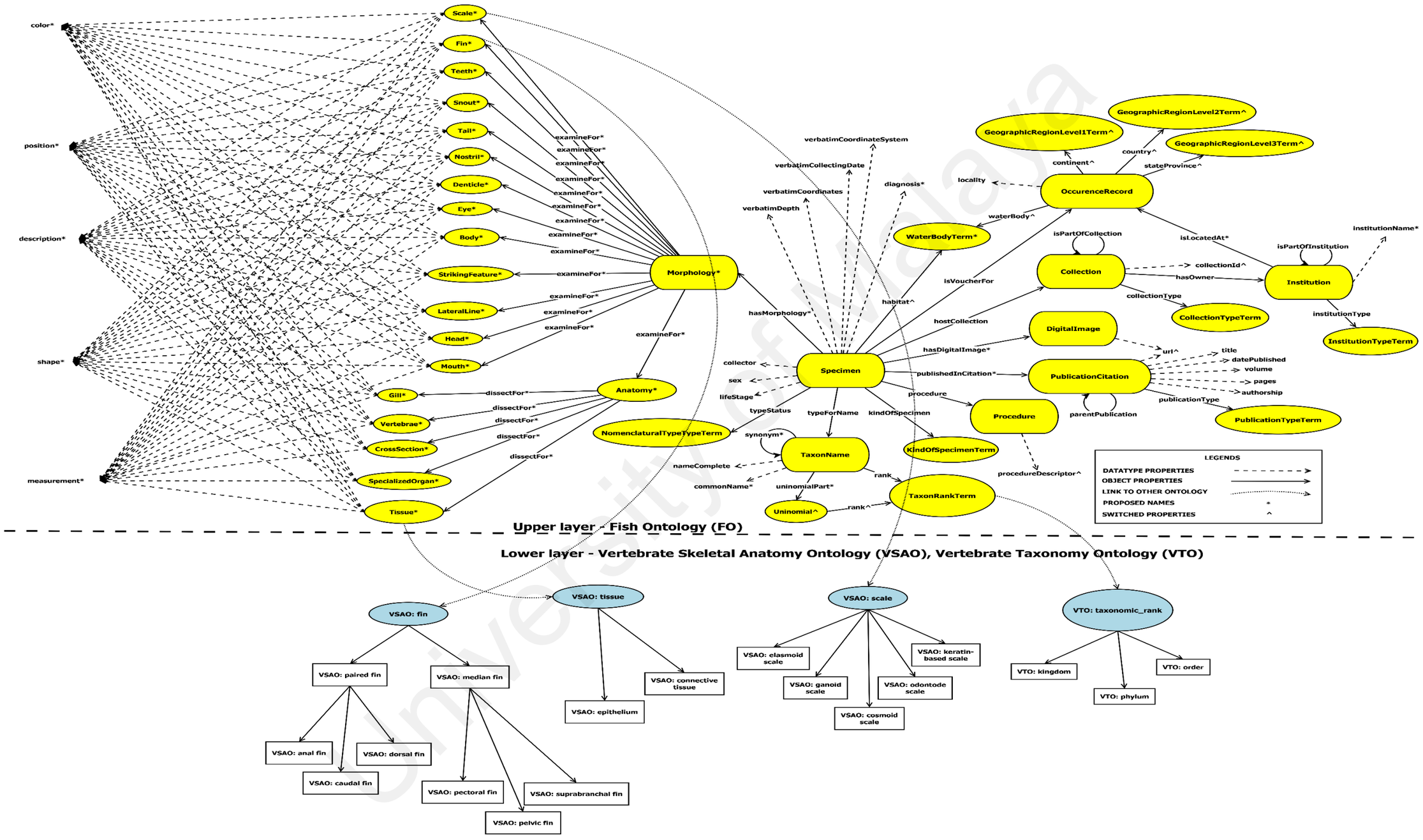


Figure 5.3: Third version of Fish Ontology (V3).

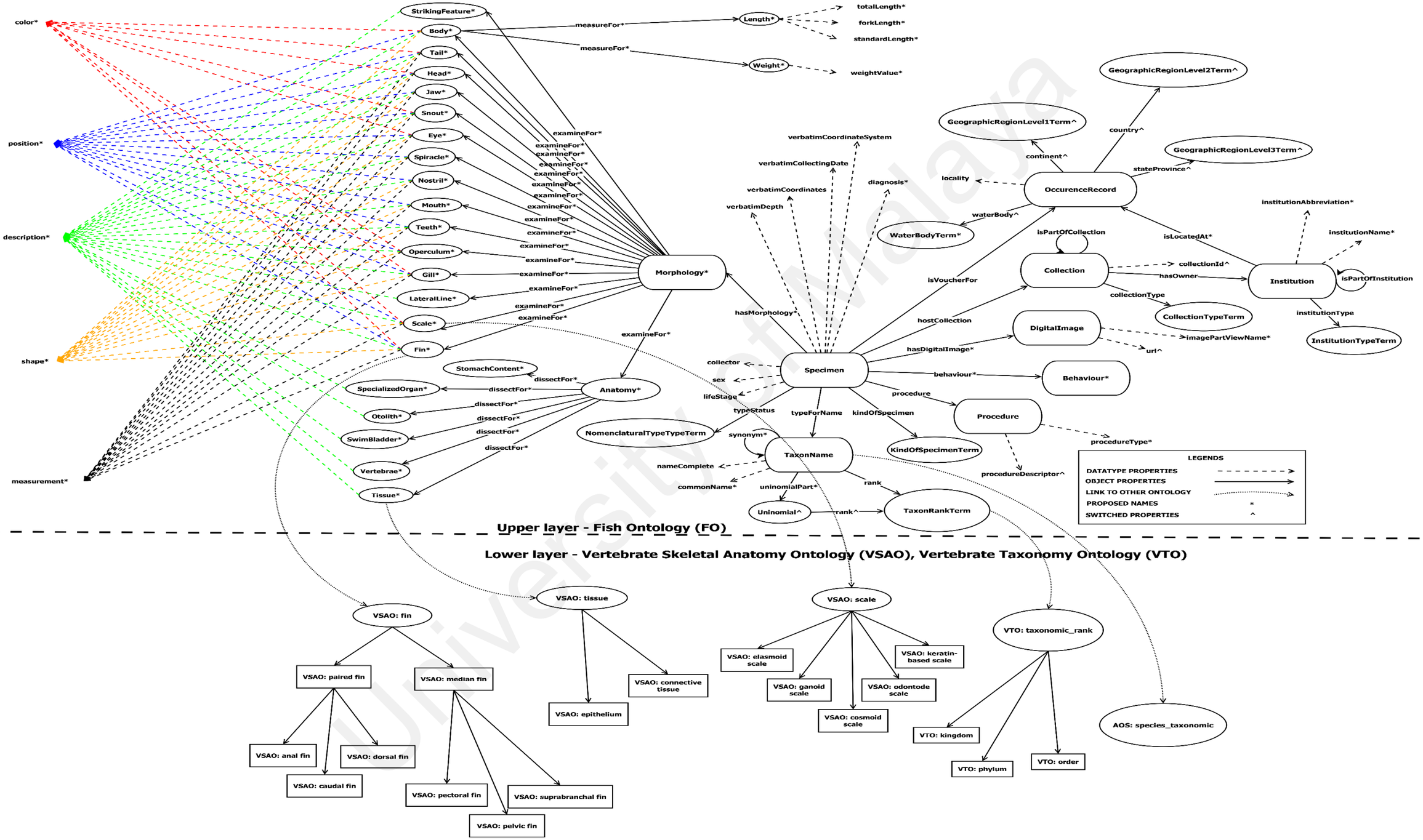


Figure 5.4: Fourth version of Fish Ontology (V4).

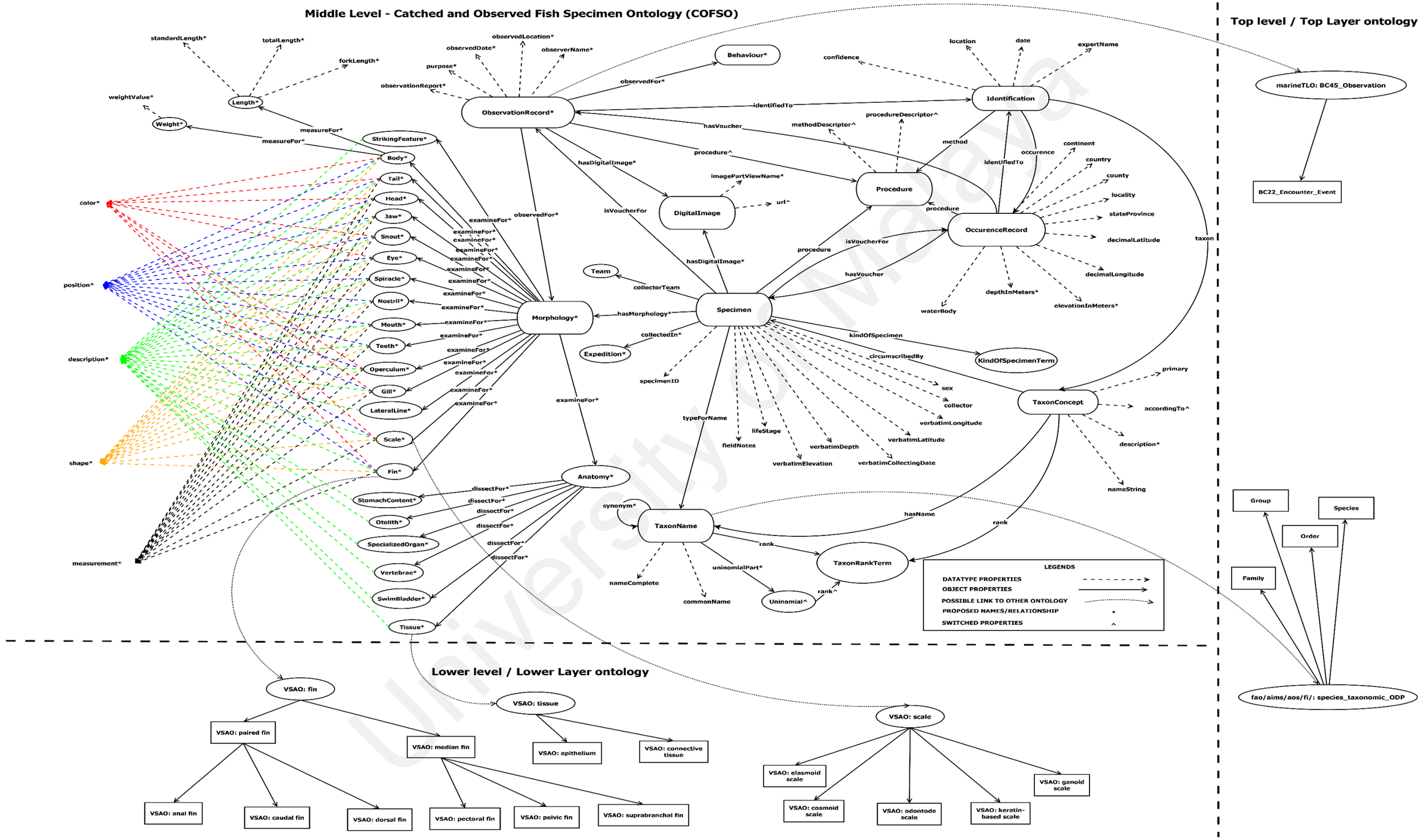


Figure 5.5: Current version of Fish Ontology

Figure 5.5 shows the current version of the FO structure. This version changes the previous ontology version from focusing on species-based information to specimen-based information in order for it to capture any specimen or sample information while still retaining its previous terms. Furthermore, in future enhancement, we would like to use it for fish automated recognition. A survey was conducted to capture the user need and awareness to enhance the capabilities of the second version of the FO and to apply some of the results of the survey to enhance the user experience of the prototype web portal. The third version of the ontology is the combination of the features from the first version and the second version of the FO. It can cater species-based and specimen-based information and has additional function to infer more results from the data that are provided in both of the ontology versions.

We have envisioned practical cases of real life applications using this ontology. As shown in the results, the FO can infer conservation and evolutionary statuses of a fish as well as show related characteristics, e.g. early jawed fish, which are useful information for interested museum visitors. The FO's ability to infer location and habitat of the fish can be useful for students or researchers. They can use the FO to identify species using local names since all fish names in the FO are linked to other database repositories. Linkage of the FO to other ontologies via reusing of terms allows the search for relevant information such as genetic data of a specific fish species. In this way, the FO is able to produce new knowledge which is useful to biologists.

In the future, we hope that FO can automatically recognize species based on the shape or characteristic provided by any specimen or sample. We hope to develop a system that can link the FO to other related portal and automatically recognizes the fish based on captured images and infer new information based on the images.

5.4 Further enhancement plan

To achieve our future vision for the Fish Ontology we need to include several enhancements to the ontology. The first imperative enhancement plan for the ontology is to complete the categories of fish for automated classification. This enhancement ensures it can recognize all the current known species of fish in the ontology. So far, the FO still has not covered all the known fish information, although more than a thousand terms have been added for the sake of fish taxonomy classification. More data need to be added to the FO in order to make it fully recognized fish species based on taxonomy. The development for classifications of several highly diverse groups, such as bony fishes, advanced jawed fish, sharks, skates, and rays, are still ongoing.

The second enhancement plan for the FO is to integrate it with the fish recognition program. For a proper future integration, the ontology must recognize the feature of the fish such as its anatomical, meristic, molecular and morphometric characteristics. We have acquired the necessary data to enhance the ontology for integration purpose from the fish expert, Professor Dr. Chong Ving Ching's research. However, the ontology still has difficulties capturing most of these values properly. Hence, the ontology still cannot generate a reasonably automated data using the current specimen in the ontology. That being said the ontology does perfectly infer species taxon rank, name information, and can infer imported specimen information to a certain degree.

The last enhancement plan for the FO is to increase its granularity by adopting and integrating it with any related OBO Foundry ontology such as the Gene Ontology, and the Disease Ontology. Both have a high research value impact outside of the fish and the fisheries research domain. Furthermore, we plan to include our previous ontology, namely the Monogenean Ontology (MO), Otolith Ontology and Monogenean Haptoral Bar Image Ontology (MHBI). Adopting and integrating them will enhance the value of

the FO since it can expand its vocabulary. This will improve the search function of the FO and will provide it with links to any related information provided by other ontology such as genetic content, publication, specific body parts, or related species.

5.5 Conclusion

In conclusion, the Fish Ontology provides the platform with all the necessary terms and relationships which help integration between databases and ontology. It can do simple fish recognition based on the taxonomic data inserted into the ontology. Understanding the information provided by the fish or fisheries research publication on the web are most of the time impossible. This is because, most of the public databases will cover the same information, while the related databases for the species are available in isolation. Integration is hard and these databases cannot be linked together as one centralized information center. The FO tackles these problems and acts as a framework to build semantic web systems for data integration applied in biodiversity research in the fish and fishery domain.

REFERENCES

Abu, A., Susan, L. L. H., Sidhu, A. S., & Dhillon, S. K. (2013). Semantic representation of monogenean haptoral Bar image annotation. *BMC Bioinformatics*, *14*(1), 48.

Alroy, J., Marshall, C., & Miller, A. (2012). *The paleobiology database*. Retrieved from https://paleobiodb.org/

Altova. (2016). *Altova GmbH*. Retrieved from https://www.altova.com/

Apache Jena. (2016). *The Apache software foundation*. Retrieved from https://jena.apache.org/index.html

Ariño, A. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, *7*(2), 81–92.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Avraham, S., Tung, C. W., Ilic, K., Jaiswal, P., Kellogg, E. A., McCouch, S., … Ware, D. (2008). The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, *36*(suppl_1), D449–D454.

Bechhofer, S. (2009). OWL: Web Ontology Language. In *Encyclopedia of database systems* (pp. 2008–2009). Springer US.

Berners-Lee, T. (2009). *Linked data - design issues*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hausenblas, M., & Kim, J. G. (2015). *5-star open data*. Retrieved from http://5stardata.info/en/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data* (pp. 56). Washington, DC: Aspen Institute.

Caracciolo, C. (2007). *Revised and enhanced fisheries ontologies*. Retrieved from http://eprints.rclis.org/15654/1/Revised and enhanced fisheries ontologies.pdf

Chang, X., & Terpenny, J. (2009). Ontology-based data integration and decision support for product e-Design. *Robotics and Computer-Integrated Manufacturing*, *25*(6), 863–870.

Chong, V. C., Lee, P. K. Y., & Lau, C. M. (2010). Diversity, extinction risk and conservation of Malaysian fishes. *Journal of Fish Biology*, *76*(9), 2009–2066.

Dahdul, W. M., Lundberg, J. G., Midford, P. E., Balhoff, J. P., Lapp, H., Vision, T. J., … Mabee, P. M. (2010). The teleost anatomy ontology: anatomical representation for the genomics age. *Systematic Biology*, *59*(4), 369–383.

Deans, A. R., Yoder, M. J., & Balhoff, J. P. (2012). Time to change how we describe biodiversity. *Trends in Ecology & Evolution*, *27*(2), 78–84.

Doumeingts, G., Müller, J., Morel, G., & Vallespir, B. (2007). *Enterprise interoperability: new challenges and approaches*. London: Springer.

Duckworth, D. W., Genoways, H. H., & Rose, C. L. (1993). *Preserving natural science collections: chronicle of our environmental heritage*. Washington, DC.

Eclipse RDF4J. (2016). *Eclipse incubation*. Retrieved from http://rdf4j.org/

Federhen, S. (2011). The NCBI taxonomy database. *Nucleic Acids Research*, *40*(1), 136–143.

Federhen, S. (2016). *NCBI organismal classification - An ontology representation of the NCBI organismal taxonomy*. Retrieved from http://www.obofoundry.org/ontology/ncbitaxon.html

Frimpong, E. A., & Angermeier, P. L. (2009). FishTraits: a database of ecological and life-history traits of freshwater fishes of the United States. *Fisheries*, *34*(10), 487–495.

Froese, R., & Pauly, D. (2000). *FishBase 2000: concepts, designs and data sources*. (Vol. 1594). WorldFish.

Froese, R., & Pauly, D. (2017). *FishBase*, *version (06/2017)*. Retrieved from http://www.fishbase.org/

Gangemi, A., Fisseha, F., Keizer, J., Lehmann, J., Liang, A., Pettman, I., … Taconet, M. (2004). *A core ontology of fishery and its use in the fishery ontology service project*. Paper presented at the Workshop on Core Ontologies in Ontology Engineering, Northhampton, United Kingdom.

García-Castro, R., Gómez-Pérez, A., & Muñoz-García, O. (2008). The Semantic Web Framework: A component-based framework for the development of Semantic Web applications. In *International Workshop on Database and Expert Systems Applications, DEXA* (pp. 185–189).

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., & Wang, Z. (2014). HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning*, *53*(3), 245–269.

Golbreich, C., Horridge, M., Horrocks, I., Motik, B., & Shearer, R. (2007). OBO and OWL: Leveraging semantic Web technologies for the life sciences. *Lecture Notes in Computer Science*, *4825*, 169–182.

Great Lakes Fishery Commission. (2009). *Great lakes fish stocking database*. Retrieved from http://www.glfc.org/fishstocking/

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*(5–6), 907–928.

Hardisty, A., Roberts, D., & Biodiversity Informatics Community. (2013). A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, *13*, 16.

Helfman, G. S., Collette, B. B., Facey, D. E., & Bowen, B. W. (2009). *The diversity of fishes: biology, evolution, and ecology*. *Atlantic* (Vol. 2). John Wiley & Sons.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C., Jupp, S., … Brandt, S. (2011). *A practical guide to building OWL ontologies using Protégè 4 and CO-ODE tools*. Manchester, England: University Of Manchester Press.

Ickes, B. S., Schlifer, B., Hansen, D., Bartels, A., & Sauer, J. (2003). *Graphical fish database browser for synthesized long term resource monitoring fisheries data*. Retrieved from http://www.umesc.usgs.gov/data_library/fisheries/graphical/fish_front.html

International Game Fish Association. (2015). *Fishing world record database*. Retrieved from https://www.igfa.org/fish/fish-database.aspx

IUCN. (2016). *The international union for conservation of nature red list of threatened species*. Retrieved from www.iucnredlist.org

Kalfoglou, Y. (2009). *Cases on semantic interoperability for information systems integration: practices and applications*. Hershey, New York: Information Science Reference.

Lanace, P. (2014). The role ontology plays in big data. In *Modeling Community Blog*. Retrieved July 18, 2016, from http://blog.nomagic.com/the-role-ontology-plays-in-big-data

Last, P. R., White, W. T., Caira, J. N., Dharmadi, F., Jensen, K., Lim, A. P. K., … Yearsley, G. K. (2010). *Sharks and rays of Borneo*. Collingwood: CSIRO Publishing.

Midford, P., Balhoff, J., Dahdul, W., Kothari, C., Lapp, H., Lundberg, J., … Westerfield, M. (2010). The Teleost Taxonomy Ontology. *Nature Preceding*, 7. doi:10.1038/npre.2010.4629.1

Midford, P., Dececchi, T., Balhoff, J., Dahdul, W., Ibrahim, N., Lapp, H., … Blackburn, D. (2013). The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. *Journal of Biomedical Semantics*, *4*(1), 34.

Motik, B. (2005). KAON2 - *Ontology management for the semantic web*. Retrieved from http://kaon2.semanticweb.org/

Nelson, J. S. (2006). *Fishes of the world* (4th ed.). John Wiley & Sons.

Neon Foundation. (2016). *Neon toolkit*. Retrieved from http://neon-toolkit.org/wiki/Main_Page.html

Nicola, D. A., Missikoff, M., & Navigli, R. (2005). A Proposal for a Unified Process for Ontology Building: UPON. In *Proceedings of the 16th International Conference on Database and Expert Systems Applications* (pp. 655–664). Berlin, Heidelberg: Springer.

NIWA. (2016). *The New Zealand freshwater fish database*. Retrieved from https://www.niwa.co.nz/our-services/online-services/freshwater-fish-database

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory*, 25. Retrieved from http://liris.cnrs.fr/amille/enseignements/Ecole_Centrale/What is an ontology and why we need it.htm

Orme, E. R., Jones, A. C., & White, R. J. (2008). *LSID deployment in the catalogue of life*. Paper presented at the British International Conference on Databases, Cardiff, UK.

Page, R. D. M. (2006). Taxonomic names, metadata, and the semantic web. *Biodiversity Informatics*, *3*(0), 1–15.

Poveda-Villalón, M., Suárez-Figueroa, M. C., García-Delgado, M. Á., & Gómez-Pérez, A. (2014). OOPS! (OntOlogy Pitfall Scanner!): supporting ontology evaluation on-line. *International Journal on Semantic Web & Information Systems*, *10*(2), 7–34.

Protégé. (2016). *Stanford center for biomedical informatics research*. Retrieved from http://protege.stanford.edu/

Reynolds, D., & Shabajee, P. (2001). Semantic portals - requirements specification. In *W3.Org*. Retrieved July 18, 2016, from https://www.w3.org/2001/sw/Europe/reports/requirements_demo_2/

Sathyanarayan, S. (2004). *Profile driven instant web portal*. Retrieved from https://www.google.com/patents/US6691106

Seltmann, K., Pénzes, Z., Yoder, M., Bertone, M., & Deans, A. (2013). Utilizing descriptive statements from the biodiversity heritage library to expand the hymenoptera anatomy ontology. *PLOS ONE*, *8*(2), e55674.

Seltmann, K., Yoder, M., Miko, I., Forshage, M., Bertone, M., Agosti, D., … Deans, A. (2012). A hymenopterists' guide to the hymenoptera anatomy ontology: utility, clarification, and future directions. *Journal of Hymenoptera Research*, *27*, 67–88.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, *21*(3), 96–101.

Shao, K. T. (2001). *Fish database of Taiwan*. Retrieved from http://fishdb.sinica.edu.tw/eng/home.php

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(2), 51–53.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255.

Sprague, J., Clements, D., Conlin, T., Edwards, P., Frazer, K., Schaper, K., … Westerfield, M. (2003). The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Research*, *31*(1), 241–243.

Tirmizi, S. H., Aitken, S., Moreira, D. A., Mungall, C., Sequeda, J., Shah, N. H., & Miranker, D. P. (2011). Mapping between the OBO and OWL ontology languages. *Journal of Biomedical Semantics*, *2*(Suppl 1), S3.

Top Quadrant. (2016). *Top Braid Composer*. Retrieved from http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/

Tsarkov, D., & Horrocks, I. (2006). FaCT++ description logic reasoner: system description. *Proceedings of International Joint Conference (IJCAR)*, 292–297.

Tzitzikas, Y., Allocca, C., Bekiari, C., Marketakis, Y., Fafalios, P., Doerr, M., … Candela, L. (2013). Integrating heterogeneous and distributed information about marine species through a top level ontology. *Research Conference on Metadata and Semantic Research* (pp. 289–301).

Tzitzikas, Y., Allocca, C., Bekiari, C., Marketakis, Y., Fafalios, P., Doerr, M., … Candela, L. (2016). Unifying heterogeneous and distributed information about marine species through the top level ontology MarineTLO. *Program*, *50*(1), 16–40.

Van Slyke, C., Bradford, Y., Westerfield, M., & Haendel, M. (2014). The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *Journal of Biomedical Semantics*, *5*(1), 12.

W3C OWL Working Group. (2009). *Web ontology language*. Retrieved from https://www.w3.org/2001/sw/wiki/OWL

W3C OWL Working Group. (2012). *OWL 2 Web ontology language*. Retrieved from https://www.w3.org/TR/owl2-overview/

Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, *74*(2), 329–356. Retrieved from http://www.fishbol.org/

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE*, *7*(1), e29715.

Yoder, M., Mikó, I., Seltmann, K., Bertone, M., & Deans, A. (2010). A gross anatomy ontology for hymenoptera. *PLOS ONE*, *5*(12), e15991.

Zheng, X., Zhang, Y., & Zhong, J. (2010). An Ontology Method for Silver Carp Auto-Recognization Based on Digital Image. *International Conference on Artificial Intelligence and Computational Intelligence* (Vol. 2, pp. 331–334).

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Publication:

Ali, M. N., Khan, H. A., Then, A. Y., Chong, V. C., Gaur, M., & Dhillon, S. K. (2017). Fish Ontology Framework for Taxonomy-Based Fish Recognition. *PeerJ*, 5, e3811.

1st Paper Presentation:

Ali, M. N., Khan, H. A., Then, A. Y., Chong, V. C., & Dhillon, S. K. (2015). *Integrating existing ontologies with TDWG LSID to build a biodiversity ontology: A case study on fish*. Paper presented at the Biological Sciences Graduate Congress (BSGC), Bangkok, Thailand.

2nd Paper Presentation:

Ali, M. N., Khan, H. A., Then, A. Y., Chong, V. C., & Dhillon, S. K. (2016). *Fish Ontology framework for fish recognition based on taxonomy*. Paper presented at the International Conference on Bioinformatics (InCoB 2016), Biopolis, Singapore.

APPENDIX

(a) *Appendix A: Questionnaire for COFSO (Second version of FO)*

Importance of specimen based global data repositories (COFSO) for research purposes

This questionnaire is crucial in siding the development of Captured and Observed Fish Specimen Ontology (COFSO), a specimen based online fish database, to ensure that it is build in the right direction while having a solid concept for specimen based global data repositories, and to determine the usability of the currently build system.

Please read all questions, introductions and instruction carefully before answering. Only fill or tick the box required for your answer. Please answer the question honestly since it is crucial for getting unbiased data. Thank you and happy answering.

Part A: Personal Information and Knowledge Level

1. **Field of Study:** _____
2. **Current education level** ☐ Degree. ☐ Master. ☐ Phd. ☐ Others: _____
3. **Do you have any experience in recording fish data?** ☐ Yes ☐ No
4. **How do you usually record it?** *(May put multiple answer.)*
   - ☐ Table in paper.
   - ☐ Table in Excel.
   - ☐ Microsoft Access.
   - ☐ Microsoft SQL Server.
   - ☐ Php MySQL.
   - ☐ Other: _____
5. **Have you heard about online database before?** ☐ Yes ☐ No
6. **Do you prefer using online system to store your data?** ☐ Yes ☐ No
7. **Please check any fish database that you recognized and used before.** *(May put multiple answer.)*
   - ☐ FishBase.
   - ☐ GBIF (Global Biodiversity Information Facility).
   - ☐ iMarine.
   - ☐ IGFA Fish Species Database.
   - ☐ Catalog of Fishes
   - ☐ NZ Freshwater Fish Database (NIWA)
   - ☐ Other: _____
   - ☐ I never know any online fish database.
8. **What do you think about sharing data about your captured or observed specimen online?** *(May put multiple answer.)*
   - ☐ I'm afraid to share my data since it might be wrong.
   - ☐ I don't want to share my data since it's my work not others.
   - ☐ My data is too important to be shared.
   - ☐ I don't mind sharing my data to anyone.
   - ☐ I like to share my data but restrict the access to who can view/use it.
   - ☐ I will share my data only on request.
   - ☐ Other reason: _____
9. **In your opinion do you think it is necessary to share specimen data online?**
   - ☐ Yes.
   - ☐ No.

Part B: Grasping Concept

Introduction for Ontology and Semantic Web technology

Ontology is a knowledge base of a domain that can give meanings (semantic) to terms or objects and allows computer to understand it like human do. It is important concept in creating semantic web technologies which can linked data together to form an information network and perform a lot of complex operations. Although currently this technology performance is a bit behind the current technologies used for database, however given its potential usage it would be one the most important technology in the future. An example of a system that have incorporated this technology is Google search and Bing search.

10. **Now that you know a bit about ontology and semantic web do you agree to apply this in fish and fisheries research area?**
    - ☐ Absolutely.
    - ☐ Not really.
11. **In a scale of 1 to 5, please rate whether you agree that having a better search system for fish and fisheries area would be beneficial for research purposes?**
    Totally Disagree ☐—☐—☐—☐—☐ Totally Agree
12. **In your opinion which is the most important aspect in gathering fish data?** *(May put multiple answer.)*
    - ☐ Specimen/Species observation
    - ☐ Specimen/Species occurrence
    - ☐ Specimen/Species taxonomic information
    - ☐ Others: _____
13. **Do you think it is important to do research on fish morphological and anatomical parts?**
    - ☐ Yes.
    - ☐ No.
14. **Among this criteria, please choose which part do you think an important morphological aspect for fish research?** *(May put multiple answer and may answer all.)* ☐ Body ☐ Nostril ☐ Tail ☐ Head ☐ Jaw ☐ Eye ☐ Spiracle ☐ Mouth ☐ Teeth ☐ Operculum ☐ Gill ☐ Lateral Line ☐ Scale ☐ Fin ☐ Striking Features ☐ Others: _____
15. **Which of this fish organ do you think is important for research purposes?.)** *(May put multiple answer and may answer all.)* ☐ Otolith ☐ Stomach ☐ Tissue ☐ Gill ☐ Specialized Organ ☐ Swim Bladder ☐ Vertebrae ☐ Others: _____

Part C: System Development

Introduction for COFSO

Captured and Observed Fish Specimen Ontology (COFSO) as its name have suggested, is an online ontology system created to store captured and observed specimen data. Not only it can store our own data, it can also link our specimen data to other online data repositories which store species taxonomy data, observation record or occurence record and create a complete fish information network. This is different from other online databases out there which provide data about species only from their own repositories. COFSO is created to automatically generate species related data from a cumulative specimen data.

For question 16 and 17, please rate each and every features.

Species Occurrence and Observation records are one of the most important aspect for storing specimen based data. Currently this is the features implemented by COFSO for storing fish occurrence record. Please rate the relevance of these features

16a. Location	Not Important ☐—☐—☐—☐—☐ Important
16b. GPS position	Not Important ☐—☐—☐—☐—☐ Important
16c. Depth and Elevation	Not Important ☐—☐—☐—☐—☐ Important
16d. Water Body	Not Important ☐—☐—☐—☐—☐ Important
16d. Specimen Identified	Not Important ☐—☐—☐—☐—☐ Important

Currently this is the features implemented by COFSO for storing fish observation record. Please rate the usage of relevance features

17a. Location	Not Important ☐—☐—☐—☐—☐ Important
17b. Date	Not Important ☐—☐—☐—☐—☐ Important
17c. Purposes	Not Important ☐—☐—☐—☐—☐ Important
17d. Field Note/Report	Not Important ☐—☐—☐—☐—☐ Important
17d. Specimen Identified	Not Important ☐—☐—☐—☐—☐ Important
17e. Parts/Organs	Not Important ☐—☐—☐—☐—☐ Important
17f. Behavior	Not Important ☐—☐—☐—☐—☐ Important

18. **When you are registering for any website such as Facebook, usually you would need to enter a registration form and edit your profile. The same happens for fish. There are 2 types of form that are usually used to insert fish information. Which among these two type of form would you prefer as a way to insert specimen data?**
    - ☐ 1 page long form (Where you need to scroll down to input all data.)
    - ☐ Multiple page continous short form (Where you need to click next according to the classification.)
    - ☐ Others: _____
19. **Do you prefer using your handphone application to insert specimen data or using a web browser from your laptop/desktop?**
    - ☐ Handphone Apps.
    - ☐ Web Browser.
    - ☐ Both.
20. **There are trillions of captured fish estimated per year. Imagine each fish specimen can be viewed like your Facebook profile. Do you think it's necessary to create such a system?**
    - ☐ Absolutely. A lot of research can be done by examining each of the data.
    - ☐ I would love such a system but I hope the computers can do calculations for me.
    - ☐ Are you kidding? Why do I need to view trillions of fish profile?
    - ☐ I would just stick to viewing species profile.
    - ☐ I don't know
    - ☐ Others: _____
21. **By linking and storing specimen data we can generate a proper information network for fish. Imagine if we can generate fish species data automatically from the collection of fish specimen data. In your opinion do you think it is necessary?**
    - ☐ Absolutely. It can generate unbiased and more accurate data.
    - ☐ No. There might be issues where we unsure whether the data is inserted properly.
    - ☐ Others: _____
22. **Would you like if the system able to show the specific location of where you catch or observe the specimen?** ☐ Yes. ☐ No. ☐ I don't know.
23. **Would you like to have a system which can automatically recognize your specimen and generate their data without the need to insert it over and over again?**
    ☐ Yes. ☐ No. ☐ I don't know.

Part D: Personal Opinion and Suggestion

24. **Now that you have an idea of what COFSO is and what it could do, please give any comment, opinion, or any suggestion about the system.**
    _____
    _____
    _____
    _____
    _____
    _____
    _____
    _____
25. **Do you think this system will benefit fish and fisheries research area in the future?**
    - ☐ absolutely because: _____
    - ☐ not really because: _____

Thank you for spending some of your time for reading and answering this questionnaire. These data would be very beneficial in the development of COFSO and guiding it to be an important fish data repositories in the future.