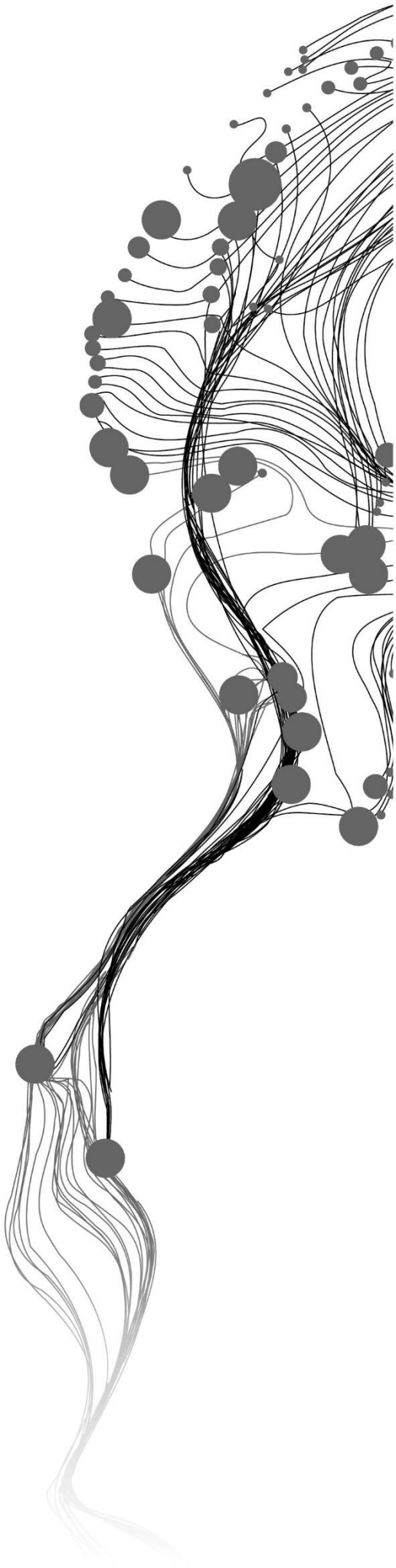

THE DESIGN AND PROTOTYPING OF AN ONTOLOGY FOR INTEGRATING CITIZEN SCIENCE DATASETS

JOSEPH YAW FRIMPONG
FEBRUARY, 2018

SUPERVISORS:
Dr. Ir. R. L. G. Lemmens
Dr. F. Ostermann

THE DESIGN AND PROTOTYPING OF AN ONTOLOGY FOR INTEGRATING CITIZEN SCIENCE DATASETS

JOSEPH YAW FRIMPONG
Enschede, The Netherlands, February, 2018

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. Ir. R.L.G. Lemmens
Dr. F. Ostermann

THESIS ASSESSMENT BOARD:
Prof. Dr. M.J. Kraak (Chair)
Dr. C. Stasch (External Examiner, 52°North Initiative for Geospatial Open Source Software GmbH)

ABSTRACT

Citizen Science is an approach to science that uses the general public in conducting scientific studies about a phenomenon or an occurrence in nature. Citizen Science makes room for the general public to measure, map and record occurrences of events on the earth's surface. These activities generate the various data and information. Most importantly, natural and environmental datasets resulting from Citizen Science projects have several qualities which can be used to increase scientific knowledge and to aid in the scientific knowledge discovery. Therefore, different efforts have been made to use such information. It is evident that potential knowledge and information can be obtained through the integration of the different datasets from the different citizen science programs. The integration of these datasets is mostly a challenge due to non-interoperability and incompatibility among the different datasets. These challenges most often come from semi-structured heterogeneous data sources. An essential requirement for Citizen Science communities appears to be a standard medium to manage the generated data and allow to integrate these datasets with other datasets for sharing and reuse. This research seeks to propose a solution for solving and managing the non-interoperability and incompatibility among Citizen Science datasets by building an ontology for data integration in Citizen Science. The design of the citizen science ontology for data integration was developed by the fusion of the IEEE standard for software development and the Generic Ontology Development Framework. The ontology was built using both spatial and non-spatial relations in Citizen Science for mapping concepts and knowledge. It was finally implemented in an OWL format. The Citizen Science ontology serves as a surrogate for structuring and modelling different datasets to have a common structure to make them compatible and interoperable. The designed ontology was used to model different datasets in Citizen Science using the Karma Data Integration tool. The modelled and combined dataset was tested using SPARQL for the different information contained in the different datasets. The results proved that the ontology is a potential tool for modelling and transforming different datasets to make them compatible with each other in the Citizen Science domain.

ACKNOWLEDGEMENTS

My most prominent appreciation goes to the Lord God Almighty for giving me life and for bringing me this far. My appreciation goes to the GNPC Oil and Gas Learning Foundation for sponsoring my education. My sincere appreciation goes to my supervisors, Dr Ir. R.L.G. Lemmens and Dr Frank Ostermann. My appreciation goes to all those who volunteered to help and encourage me, especially Brother Kwame(ICF), Brother Jacob (ITC), Brother Solomon (GNPC), Pastor Amoah (AG), Sister Tandor (UG), Brother Samson and Sister Dinah Ogara(ITC). Thank you for your support and encouragement. Finally, to my parents and family who always encouraged and prayed for me, I am so grateful.

This thesis is dedicated to My Dad
Dr Paul Kobina Frimpong
Thanks for the love and Support
"I LOVE YOU DAD"

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and Problem Statement

Citizen Science makes room for the general public to measure, map and record occurrences of events on the earth's surface. Many Citizen Science projects stem from grassroots initiatives born out of amateur hobby perspectives where the primary motivation for such communities is sharing and shared learning (Tinati et al., 2017). These activities generate huge amounts of both spatial and non-spatial data. Although there has been an enormous increase in data storage capabilities (Viswanathan et al., 2013), challenges in data sharing still cause Citizen Science communities to organise their data differently for every next project (Knights, 1976). These challenges involve data redundancy problems and also system performance. An essential requirement for Citizen Science communities appears to be a standard medium to manage the generated data and allow to integrate these datasets with other datasets for sharing and reuse. Any Citizen Science community aims to create science-based understanding through collaboration between the citizen scientists (non-professional scientists) and scientists (professionals). Researchers embedded in these communities also turn to empower the citizen scientist to ensure effective collaboration (Kolok et al., 2011). Citizen scientists are individuals with little scientific background, but are interested in contributing to discovery in science (Tinati et al., 2017). They form an integral part of groups with little or no supervision from professional scientists and undertake projects that include data collection and processing. A Citizen Science community aims to gather, organise and share the generated datasets among professional scientists and researchers with different backgrounds across the globe (Lukyanenko et al., 2016). In this project, a group of citizen scientists and Citizen Science communities that are keen to generate and use spatial data are referred to as Geo-Citizen Scientist and Geo-Citizen Science communities respectively, and all data generated is called the Geo-Citizen Science data. Spatial data are a fundamental component in decision-making for most scientific research activities (Matthews et al., 2013). In this regard, large repositories (data warehouses) and spatial data infrastructures, such as the GIS LOUNGE and INSPIRE for spatial data storage, analysis and management have evolved (Viswanathan et al., 2011). Such spatial data warehouses and infrastructures serve as the data backbone for some organisations to make spatially informed decisions (Lee et al., 2015). Many other Citizen Science projects and non-Citizen Science research activities also require different sets of spatial data for making decisions (Lee et al., 2015) and therefore could adopt data technology from previous Citizen Science community projects. This, however, does not happen due to challenges in data sharing associated with incompatible datasets, thus making data sharing and reuse in the Citizen Science community problematic. An example of the non-Citizen Science community that can benefit from Citizen Science data is the Earth System Science and Environmental Management (ESSEM), developed under the European Cooperation in Science and Technology (COST) Action program (Joffre, 2010). ESSEM addresses spatial issues under the Cost Action Program. Examples of some of the projects under ESSEM are the COST Action 719 and the COST Action ES1308. The COST Action 719 is a completed project that aimed at improving the use of geographic information system(GIS) to increase scientific research in the fields of climatology and meteorology (COST, 2011). COST Action 719 project uses various datasets concerning climate and weather, organised from different platforms to analyse the operational use of GIS applications and databases (Oussous et al., 2017). However, as in the data structuring challenges faced by Citizen Science communities, this project also suffered from problems of data sharing and reuse due to lack of standard and concise procedures for integrating heterogeneous datasets. The final report suggested that

sharing and integration of climate-related data in different digital formats is relevant but is hindered by the lack of common standards (Dröes et al., 2009). The second ESSEM project, ES1308, is an ongoing program that aims to develop a network for the climate change research communities to provide solutions to climate-related challenges (COST, 2013). The project has a fundamental component in the sharing of climate-related data among researchers and modellers. In this regard, Citizen Science data from both ongoing and existing projects can be adapted to facilitate the entire project. Old data sources sharing, and reuse requires a proper medium for operationalisation and dissemination. The challenges in this area are in semantic heterogeneity, non-interoperability, syntactic heterogeneity, language barrier and uncurated data (Pettibone et al., 2016). Semantic and syntactic heterogeneity is the structural and relationship differences that exist in datasets that originate from different platforms (George, 2005). These differences cause dataset merging difficulties. Non-interoperability is the inability to exchange data between systems and use it without difficulty (Da Silva et al., 2006). One of the challenges with non-interoperability is the difference in formats and applied standards, sometimes caused by different modes of collection (Frechtling, 2002). Since most Citizen Science projects occur in particular places with unique location characteristics, those local to the area may become the next citizen scientist for the project, possibly causing language differences. This appears to happen in many Citizen Science projects (Lukyanenko et al., 2016). Language barriers also contribute to heterogeneity problems of the datasets. Based on data needs and project purpose and/or protocol, upcoming Citizen Science projects can adapt to reuse relevant previous Citizen Science data since they are relatively useful. Nonetheless, the issue of sharing and reusing of such data always requires integrated and standardised datasets for proper dissemination and use.

1.2. Research Identification

The problem of non-interoperability occurs in many different domains; this has caused for thorough research for a means of dealing with the non-interoperability issues. From literature, the use of ontologies has been proven to be one of the most common and best practices for solving non-interoperability issues. Ontology design is the art and science of defining a set of concepts in a domain that describes the properties and the relations that exist among them for data integration (Staab et al., 2007). Ontology design is one of the suitable ways to solve non-interoperability and data sharing issues. In this regard, this project seeks to understand how Citizen Science ontology can be designed and implemented in Citizen Science communities. The research will emphasise on building an ontology for the Citizen Science community that will support data integration and data sharing. It will finally create an integrated citizen-generated geodata prototype to demonstrate the integration with existing geodata using the ontology.

1.2.1. Research Objective

This research aims to build a Citizen Science ontology using spatial and non-spatial relations and concepts in Citizen Science for data integration and sharing.

Specific Objectives

The following are the specific objectives that will help in accomplishing the main objective of the research.

1. To develop criteria informed by literature on the design of Citizen Science ontology to select an appropriate method.
2. To build and implement a Citizen Science ontology for Citizen Science projects and datasets.
3. To assess the quality of the Citizen Science ontology.
4. To deploy the designed Citizen Science ontology.

Research Questions

Specific objective one

I. What are the criteria for selecting a methodology for the design of the ontology?
II. What are the key components and the principal requirements for ontology design?
III. How can the principles behind ontologies be applied to concepts in the Citizen Science domain?

Specific objective two

I. What are the user requirements for the Citizen Science ontology?
II. What are the criteria for defining ontological classes in the Citizen Science ontology?
III. How will the relationships between classes in Citizen Science be established?
IV. What are the requirements for implementing the Citizen Science ontology?

Specific objective three

I. What are the strategies for testing the quality of the Citizen Science ontology?
II. What are the quality criteria to be used for the metric suite testing?
III. What are the strategies for integrating the Citizen Science ontology into the mainstream ontologies?

Specific objective four

I. How will the developed Citizen Science ontology be published?
II. What are the strategies for maintaining the developed Citizen Science ontology?

Research Innovation

This project proposes to design a new ontology for integrating citizen-generated geodata and existing geodata. The Citizen Science ontology will emphasise on spatial data from the Citizen Science community. This ontology will bridge the gap between the spatial and non-spatial component of Citizen Science and propose means of improving the concepts in data integration paradigm.

1.3. Overview of Project Setup

This section aims to give an overview of the steps followed in the research to answer the research questions and achieve the research objectives. Figure 1-1 shows the overall project setup which is made of six (6) stages. Each stage is structured to describe the activities conducted to achieve a research object in answering the proposed research questions. The project starts by reviewing Citizen Science, reviewing ontologies and ontology design principles, conducting a literature review on existing efforts for designing ontologies. The Citizen Science ontology is designed based on the findings from the three stages. The next stages consist of implementation, quality testing and deployment of the ontology for the Citizen Science community. An abstract of each stage in the six-stage process are as follows:

Step 1: Review of Citizen Science: This stage in the research phase introduces the concept and practice of Citizen Science and provides information on aspects relevant to the MSc research, e.g. tools, projects, and data characteristics, data usage and data structure. The objective of this stage is to provide an adequate understanding of some aspect of Citizen Science. These aspects serve as the foundation for the design of the Citizen Science ontology. Its present facts and potential information of the domain by considering the necessary projects with potential information for developing the ontology. Citizen Science is considered as an approach to science, where different information and knowledge are obtained from the general public. The use of the information and data from Citizen Science approach requires a thorough understanding of

both the projects that generated the data and the data itself. The information and knowledge acquired from this stage enhance the understanding of selecting and reviewing specific knowledge and concepts for the design of the Citizen Science ontology. The details of this stage are expressed in Chapter Two of the thesis work.

Step 2: Review of Ontologies and Ontology Design: Ontologies are used in almost all aspect of computer science especially in artificial intelligence as a means of representing knowledge and information on the semantic web. This section aims at giving a general overview of ontologies, the characteristics of ontologies and criteria required to build a functional ontology. It also reviews some of the proposed methodologies for building suitable ontologies. The potential criteria are applied to the selected methodology to select an appropriate method for designing the Citizen Science ontology. After the review, none of the selected methodologies contains all the selected criteria. However, the Generic Ontology Development Framework stands out to be an adequate methodology for the design of the Citizen Science ontology. The section concludes with different efforts made in developing ontologies for different communities. The selection of the communities and efforts are tuned to have direct links to the Citizen Science domain. The review of the relevant work groups the existing efforts into Spatial ontologies and non-spatial ontologies. This stage is thoroughly discussed in Chapter Three of this report. The selection of the different efforts leads to the selection of different frameworks and tools for designing and testing the ontology. The next section discusses the selected framework and other frameworks used to design the Citizen Science ontology.

Step 3: Frameworks and Use Cases: This stage aims at giving a general overview of the Framework and languages used for the design of the ontology. The IEEE standard for Software Development Life Cycle merged with the Generic Ontology Development Framework are the frameworks considered at this stage. The stage discusses a general notion of spatial relations that were adopted and used during the design of the ontology. It finally presents possible use cases to serve as proof of concepts and to test the quality of the Envisaged Citizen Science ontology. This stage forms the fourth Chapter of this thesis report. Moreover, it informs the design of the citizen science ontology at the design and implementation stage.

Step 4: Designing and Implanting the Ontology: The design of the ontology is based on the information obtained from the three previous stages. This section aims at describing the process used for designing the ontology. The design starts with ontology management activities where the domain of citizen science is conceptualised into different upper-level concepts to give logic and consistency to the domain concepts. The selection of the upper-level concepts is based on the relevant aspects of Citizen Science discussed in Chapter Two. The conceptualised upper-level concepts were used in the development section (Generic Ontology Development Framework) for acquiring the different information and different concepts. The granularity of concepts was based on the upper-level concepts formulations. Subdomain scopes are defined using the semantics of the information and datasets acquired in the knowledge acquisition section. The formularisation of the ontology was performed based on a proposed semantic structure of the Generic Ontology Development Framework. The ontology concepts are expressed in the OWL 2 formal ontology language. Moreover, the ontology is edited and validated in the protégé ontology editor with the HelmiT reasoner as an implementation strategy. The designed ontology is then used for modelling different datasets in the Citizen Science domain for data integration in the next stage.

Step 5: Quality Testing: The act of using an ontology as a surrogate for the semantic in a domain has never been natural in ontological engineering. However, the Citizen Science ontology is designed to enable two or more systems (datasets) to be compatible with each other and to increase the sharing of such

information on the semantic web. This section describes data modelling using the ontology; quality testing strategies serve as proof of concepts, ontology maintenance and ontology deployment strategies for the designed ontology. The ontology was imported into the Karma Data Integration Tool for modelling different datasets. The data are modelled to test for the capability of the ontology by making the different dataset compatible with each other. The results of modelled datasets are published in RDF triples. A list of SPARQL queries developed from a set of competency questions is used to inquire different information from the different dataset in the RDF triples. The final result is presented in a table. The next stage discusses the reflection on the results obtained.

Step 6: Discussion and Conclusion: This stage serves as the final stage in the thesis work. It aims to discuss the general overview of the criteria and methodology selection, the design and implementation of the ontology and most importantly the quality testing results obtained. If finally concludes by reflecting on the design process and attempts to give general answers to the research questions.

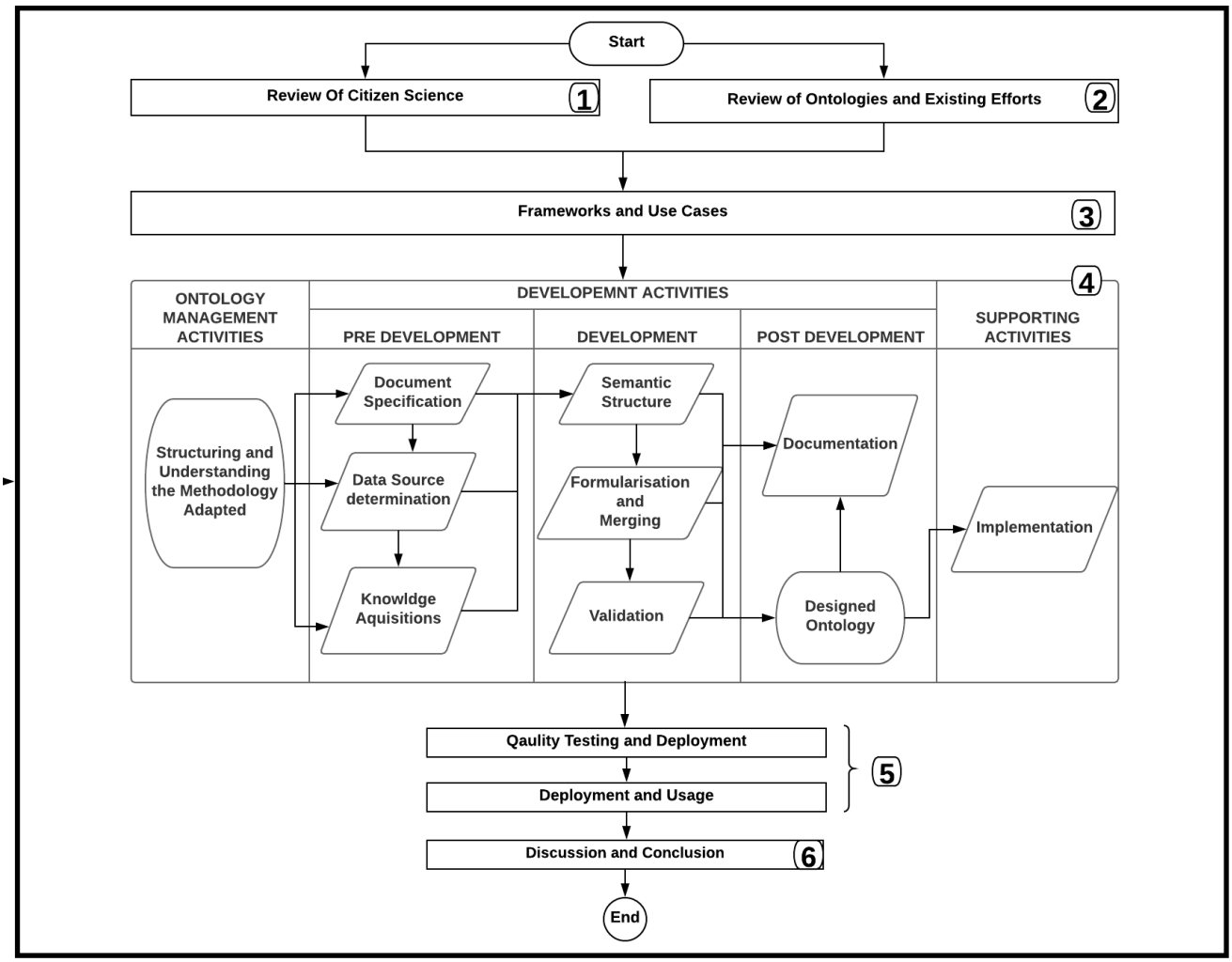

Figure 1-1: Over Project Setup. Source: Author

2. CITIZEN SCIENCE

This chapter introduces the concept and practice of Citizen Science and provides information on aspects relevant to the MSc research, e.g. tools, data sources(projects), and data characteristics.

2.1. What is Citizen Science

There have been several efforts to define Citizen Science using both its traditional and modern capabilities (Socientize, 2013). The term was first introduced by Alan Irwin in his book "Citizen Science: the study of people" (Irwin, 1995). Some of these efforts include the European union green book on Citizen Science which defines Citizen Science as the overall citizen's engagement in scientific research activities, where people effectively add to science either with their intellectual efforts or encompassing information or with their tools and resources (Socientize, 2013). Hand (2010) also defined Citizen Science in a general broad term as public participation in research. Most literature on Citizen Science does not agree on an exact date when Citizen Science was realised and used. Citizen Science activities were conceived decades ago, and it was entirely realised to have started in the early 1900s, where the general public massively contributed to data collection on locust invasion in China (Tinati et al., 2017). A general history of Citizen Science developed in an ecological study by Miller-Rushing et al (2012) records that, the use of Citizen Science in ecological studies were comprehended in centuries. Apart from Citizen Science, other exciting networks such as social media, which includes Twiter, Facebook and others allude to the development of scientific observations and aims to contribute to scientific discovery. Among these eye-catching activities, Citizen Science stands out to be one of the most exciting and innovating approaches to science which allows the public to massively contribute to scientific activities to help solve challenging scientific problems (Jollymore et al., 2017). The activities of Citizen Science range from environmental science to natural science (Sullivan et al., 2014), thereby increasing the scope and diversity of this domain. The diverse range gives the term no restrictions on any specific scope or study area. However, the term is widely used mostly in the biodiversity domain. Hence, the biodiversity domain defines Citizen Science as volunteer's collection of biodiversity and environmental data for increasing the knowledge about the natural environment in the scientific world (Pettibone et al., 2016). Citizen Science has other aliases that describe the activities condone with this approach to science due to its diversity (Eitzel et al., 2017). Most of these aliases try to capture the intended purpose and put Citizen Science in a specific context. Some of the aliases include amateur science (Bonney et al., 2009), crowdsourced science (Hoedjes, 2014), volunteer monitoring (Ledermann et al., 2015), Volunteered Geographic Information (Jollymore et al., 2017), neogeography (Turner, 2006), public participation (Pavlic et al., 2013) etc. All these aliases make the domain of Citizen Science broad and diverse thereby increasing the data collected and other activities from the public. The powerful capabilities of Citizen Science promote the term big data by contributing a chunk of citizen-generated data for scientific work. A lot of the collected data about the natural environment from the public give the public exercising and enjoyable activities to be done in their leisure time. Scientist and researcher also get the necessary data needed to make valuable scientific decisions. These privileges given to both parties (scientist and the public) has increased participant desire and understating to take part in scientific work to produce scientific findings and to help solve challenging scientific problems. Most participants in a Citizen Science projects are born out of curiosity, thereby defining the alias amateur science (Tinati et al., 2017). The literal meaning of amateur science clarifies the fact that the public participants in Citizen Science may or may not be professional scientists, but most have the motivation of developing knowledge together with scientists (Irwin et al., 2003). The motivations from the public and the participating scientists have led to a level of increasing scientific knowledge in many academic areas. As the public observe natural occurrences, record, and share the data

obtained from these phenomena, scientists gain more understanding of nature and the world. Moreover, these contributions from the public help answer some scientific inquiries and essential scientific questions. Researchers driving Citizen Science ventures are keen on the consistent yields and the outcomes created from the datasets for legitimate scientific work (Socientize, 2013). The ultimate aim of scientific work is to understand the environment and to improve the quality of life for citizens. Most of the results obtained from scientific work are directly meant for citizens. Therefore, the results are much appreciated, accepted and understood if, the citizens themselves are fully involved in the scientific processes (Johnson et al., 2014). Involving the general public thoroughly in a scientific project may require a clear and concise standard for clear communication to help improve the credibility of the information obtained from such projects Developing a standard means for managing data and information will yield an efficient result when a clear scope is defined for the project and data gathering activities (Pocock et al., 2014). Citizen Science projects are no doubt to benefit from such advantages. Hence, a clear standard can improve communication among the parties involved in Citizen Science projects. However, Burgess et al (2017), concludes that a Citizen Science project is efficiently communicated if the location of the project is considered in the analysis stage. Since projects occur in a particular place over a period, the locals in such areas become the direct participants in the projects thereby influencing the datasets. This in effect affects the knowledge and the intended purpose of the Citizen Science project. Almost all Citizen Science projects consider the location of the projects. Therefore, there is always an aspect of geospatial information associated with the Citizen Science projects as well as the data and information derived from Citizen Science projects.

2.2. Geospatial Component of Citizen Science

This section describes the geospatial component of Citizen Science, hence the term Geo-Citizen Science. Citizen Science can provide a large volume of data across a diverse range of locations and habitats over time (Bonney et al., 2009). These significant variations in location provide spatial information that comes along with almost all Citizen Science projects and data (Bhattacharjee, 2005). However, some projects are solely designed for geospatial information (Zorica et al., 2010). The geospatial component ranges from both the location of projects, the spatial attribute that comes along with the citizen-generated datasets and other spatial relations involved in Citizen Science (Coleman et al., 2009). The spatial information in Citizen Science attaches meaning to the datasets for easy and efficient analysis to inform spatial decisions. Spatial data are crucial in decision making in our day to day activities. From a research perspective, spatial data for analysis forms the basis for both demographic and geographic or spatially informed decisions in the scientific world (Mäkelä, 2006). Therefore, spatial data generated from Citizen Science contributes to spatially informed decisions in the scientific world. The importance of spatial data calls for the need to examine the characteristics of spatial data from projects and resources with Citizen Science inclusive. A general overview of spatial data as proposed by Kraak & Ormeling (2011) is said to compose of three most important sections. These components include the location, the attributes/metadata and the temporal issues. These three significant characteristics of spatial data help in answering the question where, how and when a phenomenon occurred. Most Citizen Science projects that focus on spatial information are referred to as volunteered geographic information (VGI) (Ballatore et al., 2013; Goodchild, 2007). Goodchild coined the term VGI which is an alias of Citizen Science in 2007 (Goodchild, 2007) to describe spatial information generated by the public. Most Citizen Science data have a geographic section, therefore, ensuring compatibility of different data from different platforms among Citizen Science community to enhance sharing and reuse of these data is keen. Many Citizen Science datasets from projects and for projects are correctly thought of as sample observations from a given universe of discourse, selected from observations that could be given a geographic location. Most of these datasets have significant variation due to the different mode of collection. It is therefore essential to be aware of the nature of variation and to analyse,

the results obtained from different Citizen Science projects for proper integration. In recent years, there have been several efforts to aggregate and compile spatial information for decision making using citizen-generated datasets. Examples of such efforts include the Geowiki and OpenStreetMap (Nov et al., 2011). Details of these projects are discussed under Citizen Science project section. Spatial data generated on these platforms are organised and managed to contribute to spatially informed decision in the research environment. There are quite some agencies, platforms and organisations that seek to govern and provide frameworks for the practice and management of spatial data. Examples are the Open Geospatial Consortium (OGC), the International Organization for Standardization (ISO) and others. The OGC in their recommendation "Spatial data on the web" reviews the characteristics of geographic data and suggests a range of guiding principles for working with the spatial data both on the web and in other applications (OGC, 1999). The ISO standard for geographic information provides a list of standards which regulate and provides guidelines for managing spatial information (Tom & Roswell, 2009). McGarigal (2001) also suggested that another essential property of geographic information is the level of detail that is apparent at scales of analysis. Meentemeyer (1989) propose a solid theoretical foundation for understanding scale when building geographic events and representations. Policies and terms of use often govern projects that describe the scope of concepts.

2.3. Citizen Science Policies and Publications

Policies in Citizen Science are sets of planned schemes of values and principles designed to guide the conduct and Citizen Science project activities (Legrand et al., 2016). Policies in Citizen Science range from all forms of legal backings and legal reforms that promote and guide the condone of the general public concerning their contribution to science. Different types of policies exist in different forms depending project's structure. These differences in policy structures between Citizen Science turns to support different forms of participation as discussed by Tinati et al (2017). Polices are mostly grouped into classifications and forms of participation regarding projects activities. Examples of such policies groupings are policies on data, policies on frameworks, policies on conducts and policies on participants. However, polices on data as well as polices on frameworks are relevant to this project. Frameworks policies in Citizen Science are mostly designed by governmental and international bodies that see to the welfare of the Citizen Science community. Examples of such bodies include the United Kingdom Environmental Framework (Geoghegan et al., 2016) and the European Commission (Figueiredo et al., 2016). They developed schemes and structure that guides the general practice of the community. Frameworks designed by such agencies serve as the building blocks for both starting and ensuring continuity with Citizen Science activities. Such frameworks also empower policies on data and information. Policies on data are mostly considered as a collection of guiding principles that regulate how a particular Citizen Science community can intermingle with other community or within a single community. Examples of these interactions may be among projects, user, Citizen Science volunteer and other parties involved in Citizen Science. Policies on data turn out to be the most well-recognised form of policy in the Citizen Science community (Bowser et al., 2013). It describes details on both the collection and use of citizen science datasets. Different kinds of policies on data are regularly found on projects' web pages. Such policies range from user agreement, terms of use, legal and private policies (Bowser et al., 2013). Most often, the different range is placed on a different section on the project platforms depending on the requirement and regulations of the parties involved. User agreement policies: they are policies that guide project that uses the information on participating members for other analysis. These types of policies are mostly found on projects' websites and must be agreed before the participant can fully partake in the project (Bowser et al., 2013). They often called "Terms Of Use" by some individual projects. However, different project recognises "Term Of Use" as the description of how information can be used. Such information includes ownership or copyright issues for the different part of the information on the platform. Legal

policies on data are often the underlining concept schemes that are within a jurisdiction governing the use of data in that jurisdiction (Bowser et al., 2013). They are mostly structured to govern both citizen science projects and its participating individuals. Privacy policies describe the management and distribution of information on data t and from the participating members. Most policies are often referred to a link the main document, or they are most often written up in documents and publications in the Citizen Science community. Publications in Citizen Science are mostly written documentation of findings and results drawn from Citizen Science projects. Publications are mostly found in journals and other scientific writing. They mostly describe the activities performs in Citizen Science. These activities range from starting a project, the methods used, results obtained and conclusion and deductions drawn from the projects.

2.4. Citizen Science Projects

There are several Citizen Science projects that have evolved and made a remarkable success in improving knowledge in the scientific world. Three examples are "ZomBee Watch" (AWS, 2010) This Citizen Science project aims to understand where in San Francisco are bees affected by zombie flies. The study helps scientist understand the spread of zombie flies in the affected areas. The "Drug discovery from your soil" (SciStarter, 2017), this project helps discover natural products (drugs) from soil fungi. It helps in identifying certain types of soil fungi that can be useful for treating certain types of disease. "Did You Feel It? (DYFI)" (USGS, 2010), this Citizen Science project collects evidence from the public who sensed or felt an earthquake instance. The information generated is used to create maps showing the experiences of the people and the extent of earthquake damage. These projects for the earthquake summary are compiled to make evacuation and other potential safety and necessary activities. Table 2-1 shows a review of some selected projects. These selected projects were based on their popularity in the domain of Citizen Science; easy discoverability and access to the resource generated from the project in question. These activities of Citizen Science projects have many components, these components range from data collection, data organising, management, designing of application to support data, just to mention a few. It is therefore vital to understand the characteristics of Citizen Science projects as well as the requirement that informs Citizen Science projects. These requirements and characteristics of Citizen Science projects inform the method and procedure to be followed during the data collection stage. Characteristics of Citizen Science can be grouped into many different perspectives depending on the purpose of such categorisation. Other means of defining the characteristics of Citizen Science are on the required datasets and protocols to follow. As project protocols may require, there are different strategies for organising each dataset for each project. Therefore, there is no unique laid down procedure to go about Citizen Science projects, more particularly data collection. However, there are many standards generated from many projects resulting in different data and data structures.

Table 2-1: Some Reviewed Citizen Science Projects. Source: Author

Project	Purpose	Location	Source
BudBurst	The BudBurst project aims at fostering collaboration among gardeners, scout-troops, climbers, botanists, environmentalists, government offices, and teachers to screen environmental change and its effects on plants. It also tries to get citizens on the field to observe how plants change with the seasons. It has over 1000 participant contributing the data collection. It is	USA	(Bryan et al. 2017) (Link)[1]

[1] http://budburst.org/

	among the most prominent projects which make data and information freely available to the public.		
GeoWiki	The GeoWiki Platform provides means of addressing global land use issues. It is an ongoing project with more than three sub-projects. Each project aims at addressing a specific global land cover problem involving the public	Global	(Link)[2]
BugGuide	The Bug Guide Program is an online community consisting of numerous naturalists with interest in observing the behaviour of insect and other species of insects. The overall aim is to organise resources concerning different species to create a knowledgebase system for all interests in insects and other species. Most of the members are in-house expertise of scientists and few amateur scientists who help to organise and collect information and identify a diversity of bug species	USA and Canada.	(Bud Guide, 2013) (Link)[3]
FrogWatch	FrogWatch was set up to attempt to and assemble better data about the frogs of North Australia and their dispersion. The program aims to provide a real awareness on frogs to the general public. This program allows groups and individuals to learn about wetlands in their communities by reporting the mating calls of local frogs and toads	Australia	(Link)[4]
Zooniverse	The Zooniverse project consists of numerous programs that promote people-powered research. This project gives volunteers the opportunity to contribute to numerous scientific research. An example is the Galaxy Zoo Project. It aims at converting citizens efforts to potentially valuable assets.	Global	(Lintott, 2017) (link) [5]
Nature Watch	Nature Watch is a Citizen Science project that provides excitement, simple and easy to use platform for ecological and environmental studies in Canada. It is organised into five modules which include the Plant Watch, Frog Watch, Ice Watch, Worm Watch and Milkweed Watch. The platform and each module are structured to provide urging a new structure for solutions to problems through data collection for researchers to use.	Canada	(Link)[6]

There are many characteristics and steps to be followed when designing a Citizen Science project; these steps inform the quality of the data generated. Alam & Gühl, (2016) concludes that there is the need to be more explicit in defining the work goals of an amateur scientist by bodies involved in the project. Besides, the United Kingdom environmental framework for Citizen Science sets some modest steps to be followed in defining a complete Citizen Science (Pocock et al., 2014). Additionally, these quality indicators can also serve as the basis for designing any system that operates on information from the Citizen Science community. The following requirements and characteristics are enshrined in Citizen Science: Why a Citizen Science project; Knowing the capability of Citizen Science can help structure the projects under consideration. This is by finding the required research question for the project. As discussed by Robertson

[2] https://www.geo-wiki.org/
[3] https://bugguide.net/node/view/15740
[4] http://www.frogwatch.org.au/index.cfm?action=cms.page&section=3
[5] https://www.zooniverse.org/projects
[6] https://www.naturewatch.ca/

(2015), if a project is explained well to parties involve, entities can cooperate efficiently to ensure well-structured and well-done project (research). Also, an excellent way to explain issues to the public is to have the well-formulated scope of the concept. This scope can be well achieved if one has clear and concise questions about the project.

a. ***Proper Research Formulation and Scope***: A suitable research problem formulated can be said to act as the foundation of a project structure. The foundations support the design of the structure. A good foundation implies a strong structure and vice-versa (Follett et al., 2015). A simple scenario is in pollution studies with citizen scientist, where Citizen Science projects seek to monitor the air quality in one's location (Kaufman et al., 2016). The public is given a set of instruments to make some recordings on the quality of air in one's neighbourhood. These recordings can be better explained if citizens know what they are measuring. This implies that scope definition can always help in facilitating operations in Citizen Science project, in this regard, defining project scope is a useful criterion of any Citizen Science project. A proper scope formulation can increase logical learning, raise individual's consciousness of their condition and enable proficient individuals to share their abilities and information.
b. ***Understanding of Project by Participating Members:*** A clear understanding of the project is a good requirement for Citizen Science project. The type of approach to be adopted to involve volunteers are crucial when explaining a Citizen Science project to participating members (Robertson, 2015). Identifying and defining project team are also a critical component of Citizen Science.
c. ***The Quality and Quantity of data:*** This depends on the number of participants. The number of participant and willingness of individuals to contribute to a Citizen Science project can be correlated unswervingly to the amount and quality of datasets.
d. ***Methods of Collection:*** Different models exist for defining the methodology for involving the public in scientific research. These methodologies include, but not limited to designing surveys, data requirements, technological requirements, storage, analysis, testing and documentation.
e. ***Data Requirement and Surveys:*** data requirement is very crucial for every project. Data from the project must be of a specific standard and quality. Therefore, plans and protocols must be followed to ensure these happen. The technological requirement also forms a reasonable basis for the type of data to be collected. Since different technologies and equipment are required on most Citizen Science projects, it is important to consider it as a requirement (Bonney et al., 2009). This requirement may include formal training and other forms of training to ensure a clear understanding of the project by the participant. This would improve the quality of the generated datasets and information obtained.

2.5. Citizen Science Data and Tools

Data collection in Citizen Science forms the basis for most Citizen Science projects. Data collection are means of organising resource to conduct a study about a phenomenon. Data forms one of the fundamental results of Citizen Science and such data are mostly organised to be used for scientific research. This section discusses the structure, and forms and data quality among citizen-generated data. The structure and forms of Citizen Science datasets serve as the basis for performing operations on these datasets and for easy parsing of such datasets over the internet by making them machine-readable.

Data Collection and Tools: Citizen Science makes use the public in collecting potential data for scientific studies. Data collection in Citizen Science uses both the traditional and modern means of acquiring information about a phenomenon (Schade et al., 2016). However, the advancement in technology recently

has increased the immense participation of the public as well as enhancing the mode of data collection. A review of the discussed project in Section 2.4 Table 2-1 shows that different tools and equipment are deployed for a different purpose depending on the project characteristics and project requirements. Tools such as websites, smartphones and sensors are widely used in acquiring data (Schnoor, 2007). These different tools have opened up different methodologies used for data collection in Citizen Science. These methodologies range from time-dependent, cost dependant to effort factor dependents among other factors. The different methodologies considered in this project are based on the tools data collection and other different factor dependants. All datasets collected with the different tools and methodologies are organised and submitted to projects platforms. After data submission, the next decisive action is the processing of the different datasets. The next stage step aims to give how accessible the submitted data are to the public ones submitted.

Data Access: Data access is often referred to as the easiness to realise a resource or datasets on a specific platform. To determine the characteristics and structure of the different datasets in Citizen Science, a description of how easy to discover such dataset was first considered. In Citizen Science, there are many enthusiasms towards collecting datasets. However, all citizen-generated datasets are not fully available to the public in their raw state on these Citizen Science platforms. Citizen Science platforms advertise and convince the public to participate fully in their research (Schade et al., 2016). They, however, don't make it easily accessible for citizens to download the dataset directly from their website. Most are released as processed and transformed information for making decisions (Roman et al., 2017). That is, the projects platforms present the outcome of the projects in the form of reports and display these reports on their website instead of the citizen-generated data. However, when a participant request dataset, it is released to them based on policies and regulations regarding the project and the use of the data. One can conclude clearly that discoverability of datasets does not mean accessibility of the datasets to the public in Citizen Science community. It can be of high benefit if access restrictions and difficult (Complex) terms of use on citizen-generated datasets can be lowered and data made available in its raw state on all platforms. This can improve the interests of the public in Citizen Science. It can also improve the quality of data during collection stage due to the high motivation that can be generated from the easy access to data. The critical aspect of data access is to provide understanding which is used in the final product.

Data Usage, Datatype and Structures: Datasets in Citizen Science have several data types and structure due to the different mode of collection. The data structure is the manner in which datasets are organised and stored in a specific format for easy accessibility and effective modification (performing operations on them) (Eitzel et al., 2017). Organizing and storing data for easy access and use are a worry in almost all field of science with Citizen Science no exception. Data structures in computer science are precisely defined as a collection of data values, data types, the relationships among data values and datatypes, and the types of functions that can operate on these values. Most Citizen Science datasets are a collection of an individual instance of sightings which portrays how an occurrence or a phenomenon occurs (mostly species in a particular biome (Biodiversity dataset)). Data types and formats, as well as data structures in the biodiversity domain, is considered these sections. A review of Citizen Science in the biodiversity domain reviews a list of data formats. Datatypes range from strings, Boolean to a natural number. These different datatypes in Citizen Science datasets allow easy manipulation of the datasets. Examples of data structures that exist include arrays, sets and linked sets, aggregate data structures (records and unions of sets). The different datatypes and structures allow different operations performed on such information to solve complex scientific problems. Citizen Science data has been used in diverse ways. Example of such usage includes quantifying spatial variation, modelling species distribution and other forms of solving challenging issues

concerning humans. Most often the results of data usage in Citizen Science is profoundly informed by the quality of the data.

Data Quality: Citizen Science community generates both quantitative and qualitative data through observation. The term data quality in Citizen Science is, therefore, a measure of how best the generated data is fit for use or its intended purpose. Citizen Science is considered as an evidence of getting the required data from projects for an envisioned purpose. However, information on data quality is an essential worry for specialists utilising and investing in scientific research based on Citizen Science approach. Evidence and literature suggest that the involvement of citizens in other forms of works other than research has a higher chance of success (Roman et al., 2017). Involving the public in research activities without care and well-implemented strategies can grade in undesirable results. There have been several measurement innovations which use ethical perception capacities to enhance data quality in Citizen Science communities. These approaches, however, have proved fatal with Citizen Science (Lukyanenko et al., 2016). From the review conducted on Citizen Science projects, there are always some trade-offs between endeavouring to get people involved in a Citizen Science project and citizen volunteering to engender datasets for a specific project when it comes to data quality. Getting quality data most often comes from people volunteering to give the datasets. Citizens volunteering themselves to produce datasets often results in highly au courant and high-quality data as compared to a scientist endeavouring to get people to contribute to projects which are not of significant interest to the citizens. Therefore, data quality in Citizen Science can be said to highly depend on how well a Citizen Science programs are structured in the execution and collection processes and how exciting projects are to participating members. However, this dataset will be challenging to merge since they were developed with different objectives. Data Integration: Interoperability is the ability for systems to easily combine different datasets or share information across different platforms with ease. When one receives or obtains different datasets from different platforms in Citizen Science, the next idea that emerges is putting together the heterogeneous datasets to make a meaningful decision from it. The integration of the datasets poses challenges due to the heterogeneity of the datasets. James Handerler argued that non-interoperability is due to the fact the datasets involved are not designed to be compatible (Hendler, 2014). Other research such as (Barbosa et al., 2014) related the problem of non-interoperability results from the difference in formats that exist. Examples of the different formats of datasets that exist in Citizen Science are Citizen Science, GeoJSON, pdf and many other. However, with the advancement in technology Citizen Science datasets with different format can easily be merged if the semantic of Citizen Science datasets are logically compatible. Computers, as well as developers, can handle heterogeneous datasets to analyse visualised and make meaning out of it if the semantics of Citizen Science dataset is clear and concise to handle. Integrating different dataset can help extract from each dataset a piece of information which when combined can solve challenging scientific problem. There has been some effort to improve the integration of heterogeneous datasets by different scholars in other domains. A renowned mean for solving non-interoperability is the use of ontologies.

3. ONTOLOGIES

This chapter introduces the concept of ontologies, the practice of ontology design, and provides information on criteria for selecting a methodology for designing ontologies. It concludes by applying the selected criteria on most frequently used ontology methodologies to select an adequate method for designing the Citizen Science ontology.

3.1. Semantic Web

The semantic web is an extension of the world wide web (www) that aims at combining a set of tools and techniques to create meaningful data on the internet (Ristoski & Paulheim, 2016). The purpose of creating semantics for data is for computers and machines to understand and use such data efficiently. In a coherent sentence, the World Wide Web Consortium (W3C) is building a stack of tools and technique to support web of data (Staab & Stuckenschmidt, 2006). These tools and technologies enable people to handle data on the web by writing rules, building vocabularies and building repositories. The idea of the semantic web is to build a smart web which can enhance communication between computers and human. Tim Berners Lee defined this ideology of semantic web in 2000 (Halpin, 2013). He proposes a structural architecture for realising the semantic web. Figure 3-1 shows the architecture proposed by Tim Berners Lee and is composed of seven layers. The first layer describes how resources are encoded and identified using a Unicode and a unique resource identifier (URI). From his proposal, all resources can be encoded and identified by URI. Encoding and identifying resource uniquely serves as the basis for the whole architecture. However, to avoid multiple naming, a unique NameSpace (NS) is given to a group of resources to avoid the collision by applications. As the second layer in the architecture, the Extensible Markup Language (XML) and Extensible Markup Language Schema (XMLs) discusses a structure and grammar for structuring resources to a common schema in addition to the namespace. The XML Schema provides a structure and XML provide a grammar or syntax for organising the resources. The third layer ensures that all the different resources can coexist. The Resource Description Framework RDF and Resource Description Framework Schema (RDFS) serves as a foundation block for different information and different resource to coexist on the semantic web. It also defines a machine-readable structure, where machines quickly understand the information encoded in the resources. The fourth layer describes the means of capturing knowledge contained in a resource or set of resources. These knowledge capturing is by defining sets of vocabularies and mapping the relationship among these vocabularies using properties and predicates. After capturing the knowledge contained in a domain, it is then essential to define the logic that exists in the knowledge captured. Therefore, the function of the logic/fifth layer is providing a reasonable understanding of the knowledge captured using the ontology. The proof/sixth layer serves as a check as to whether the results obtained from a resource search are valid as requested by a query. Lastly, the trust layer serves as a framework for simple data transfer and extending transactions. Each layer presents some essential techniques for the semantic web. However, the primary concern of this thesis is the ontology layer for capturing domain knowledge.

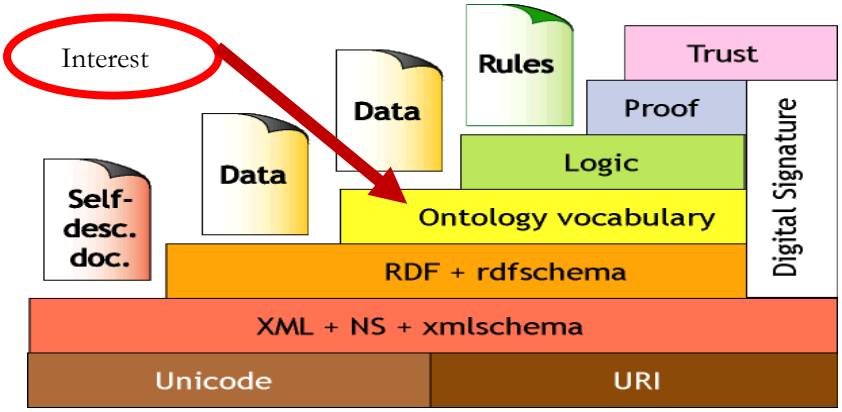

Figure 3-1: The Structure and Architecture of the Semantic Web Source: (Berners-Lee, 2000)

3.2. What are Ontologies

Ontologies were first discovered in philosophy as a branch called metaphysics, that deals with the nature of being or what is out to exist (Varzi, 2011). As it was a notion in philosophy, philosophers described ontology as a way of studying modes of being and the interactions that exist in nature or the universe (Coffey, 1938). The computer science society later adopted this idea of studying the nature of being. Ontologies were adopted as the core research interest in artificial intelligence (AI), (a branch of computer science that deals with intelligent agents). AI adapted ontologies as an appropriate means of capturing knowledge for building intelligent systems (Hadj et al., 2014). AI discusses the importance of using ontologies as a powerful computational tool for reasoning and analysing models. The term was later clearly defined by Tom Gruber in his paper "Toward Principles for the Design of Ontologies Used for Knowledge Sharing" as "a specification of a conceptualisation" (Gruber, 1995). Ontologies, as described by (Pan et al., 2009), is said to be the art and science of constructing a Conceptual Model for capturing, preserving, and sharing domain knowledge for efficient integration of knowledge in a domain. However, there have been other definitions proposed by different researchers, but, Gruber's definition stands out to be a clear and precise way of defining ontologies. Other different definitions define ontologies in the context they work and the domain they work. Moreover, in a technical report in 2003 by Nieto on ontologies, there are reviews of definitions that possibly describe ontologies base on purpose (Nieto, 2003). He concluded that all the proposed definitions of ontologies present several common characteristics compared to Gruber's definition.

These several characteristics present the notion of ontologies based on their characteristics. The following are reviewed importance and characteristics of ontologies based on their capabilities. These characteristics include:

1. **The ability to build knowledge-based systems with ontologies:** Knowledge-based system applies astute reasoning in solving problems that may require more human time, endeavour and expertise in a domain (Corsar et al., 2008). A knowledge base is a prominent group of artificial intelligence that seeks to model systems to capture and preserve information from entities (Sajja et al., 2010). The notion of ontologies is to capture the knowledge in a domain for easy sharing and reuse. Therefore, the incorporation of ontology design in knowledge base promotes natural and logical reasoning for capturing domain knowledge. Knowledge-based systems are developed on the basses of the human brain. Therefore, preserving knowledge in the form of instances and classes in ontologies allows machines (acting like human brains) and agent to operate, invoke and use the information stored in the ontologies (Guarino et al., 1995). Ontologies are structured in a way machines will understand and use (Gruber, 1995).

2. **Ontologies can exist as storehouses for organising and managing information for a domain**. A domain may contain different subdomains that can be modelled independently to capture specific knowledge. When these subdomain ontologies are merged to form the central ontology, the different knowledge stored in the main ontology serves as the repository for the sub knowledge

represented in the central knowledge. However, ontologies can operate on databases to extract any needed information based on the schema of the database and that of the ontology. Example of such usage of ontologies is the herbal medicine knowledge repository (Mustaffa et al. 2012).

3. **Ontologies for integrating heterogeneous information sources:** heterogenous information source can be integrated with the help of ontologies based on the semantics of the information (Buccella et al. 2011). Ontologies designed for digital libraries provide the benefit of integrating the resource in the library. Ontologies can be designed based on certain qualities to either integrate information resources or allow other resources to use a particular ontology. Moreover, ontologies are a means of expressing the knowledge in a domain.
4. **An ontology expresses the metadata of a given data to present the meaning of that data**. Most of the time information that arises from the geospatial domains are in the form of observation which includes images. These observations come with a metadata attribute; these attributes mostly express the meaning contained in the observations. When an ontology is deployed to extract the knowledge encoded in such metadata, ontologies can describe well the information contained in the ontology.

Considering all the above qualities of ontology, some ontologies are designed to enable knowledge integration, knowledge sharing and reuse of resources. Ontologies serve as a means of standardising semantic web content and easing the sharing of different knowledge on different domains on the web. The quality of an ontology for any domain can be assessed by considering the quality of the mapping relationships in the ontology. These mappings provide several ways of rendering and transforming different data source to provide a well-structured system (Stuckenschmidt & Visser, 1999).

The benefits of using a well-structured system with an additional interpretation strategy and the capability to retrieve evident and new knowledge are the basis for defining a functioning ontology. Enabling data reuse is an aspect that characterises data interoperability (Chenguang et al. 2015). This calls for a precise definition of terms involved in ontologies and ontology design. These terms give a clear description of instances, concepts, attributes, and relations which forms the basis for the design of ontology. Instances of objects are the general ground level knowledge that contains the concepts in a domain. Classes are a collection or set of these concepts with attributes of the classes as the properties or characteristics of these classes. Relations are the meaning that is shared between or among different classes.

Figure 3-2 aims at throwing more light on the above terms describing the benefit derived from ontologies and how the ontological terms are used in a simple ontology to model a domain. It shows how a simple ontology tries to capture and model concept in a forest biome. The forest biome is ***FoundIn*** **San-Francisco**, and consist of the SuperClasses **Birds**, **Insects** and **Forest**. These classes have subclasses which include **Owl**, **Dragonfly** and **Coniferous-Forest** respectively. Subclass **Owl** has two individuals describing the members of the class **Owl**. There are three (3) individuals in the whole ontology. The Individuals in the ontology are described with the ***"Type-Of"*** Relation. The relation between **Birds** and **Insects** is the ***FeedOn*** relation which creates the notion that some **Birds** eat **Insect**. The ***HasHabitat*** relation is used to describe the relationship between the three superclasses. Therefore, one can logically draw the inference that, if **Birds** and **Insects** have a habitat forest, they can normally be found in the same geographic location. Therefore, the ***FeedOn*** relation maybe valid since individuals of the two classes are in the same locations. This and other inferences such kind make ontology powerful and logically adequate to conclude from facts for information processing. This model in Figure 3-2 can be extended to include different species. This conceptualisation can exist at different levels and different structures resulting in deferent types of ontologies.

.

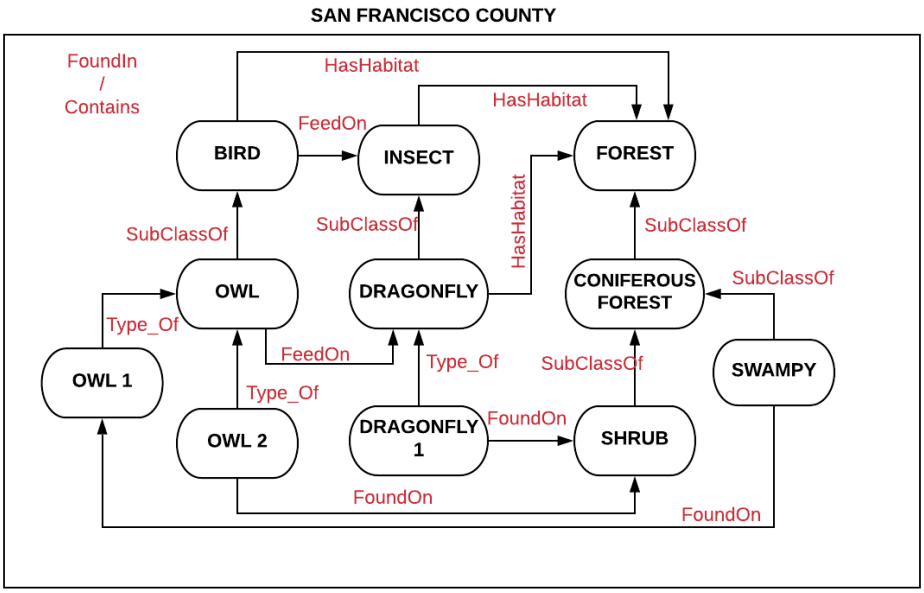

Figure 3-2: Simple Ontology Describing a Forest Habitat in San Francisco County.
Source: Author

3.1. Types of Ontologies

Ontologies are mostly classified based on the purpose they were created. A review of the types of ontologies depicts three general kinds of ontologies. However, based on the use of ontologies, several types of ontologies can exits. These three types of the ontologies considered in this section include the upper ontologies, domain ontologies and hybrid ontologies. An overview of each type of ontology and examples of already existing ontologies build with that type are discussed below.

Upper Ontologies

This type of ontology portrays incredibly intellectually-deep thoughts which are shared between different domains. An upper ontology otherwise called functional ontologies are models of the essential pieces of information that expresses different concepts for a broad range of domain (Boyce et al., 2007). Upper ontologies utilise a foundation glossary that contains the terms and related information depicted. This information mostly forms part of a universal set of keywords (Information). Institutionalized and highly accepted upper ontologies accessible for utilising include Basic Formal Ontology (BFO), Unified Foundation Ontology (UFO), Business Objects Reference Ontology (BORO), Dublin Core (DC) ontology, General Formal Ontology (GFO), The Suggested Upper Merged Ontology (SUMO), Upper Mapping and Binding Exchange Layer (UMBEL) and Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE).

Domain Ontologies

A domain means a distinct set or subset of a thing. In the case of Citizen Science, a domain can vary depending on the area of interest. A domain ontology is mostly called a domain-specific ontology. The adjective specific gives it a well-defined and distinct meaning from other domains (Oliveira et al. 2006). A domain ontology captures or models a specific concept or some concepts which belong to part of a Whole (thing). This type of ontology expresses relevant terms applied to that domain. Example of such ontologies includes the Plant Ontology (PO) and the Social Insect Behaviour Ontology (SIBO). Since domain ontologies develop concepts in precise and consistently many different approaches, they are as often as possible incompatible with each other when combined. Systems and applications that rely on domain ontologies are developed routinely need to combine the different domain ontologies into a more comprehensive ontology during operation. However, the different developing strategies do not make them

possible. Distinct ontologies in a similar domain develop due to different approaches, assorted proposed utilisation of the ontologies, and various perspective of the domain.

Hybrid Ontology

The hybrid ontology is a merged concept of both the upper ontology and the domain ontology. Hybrid ontologies are carefully structured not purposely for a specific domain, but for a specific application (Nieland et al. 2015). They are often termed as application ontologies. Hybrid ontologies when structured to enhance retrieval of information are referred to as lightweight ontology (Miller et al. 1990). No matter the type of ontology under consideration, there are laid down sets of formal languages for expressing the kind of ontologies to be built.

All the types of ontologies are governed by potential criteria that inform the quality of the final ontology as a tool for the semantic web. The next section considers the different criteria adequate for selecting a proposed methodology for the design of a quality ontology.

3.2. Criteria for Selecting an Ontology Methodology

There are many qualities exhibited by ontologies in the execution of what they were built for. These qualities are reviewed in this section. Every ontology should have specific characteristics; these characteristics serve as useful indicators or criteria for developing the ontology under consideration (Gruber, 1995). Tom-Gruber (Gruber, 1995) proposed five design criteria for building ontologies, His proposed criteria have now become an integral part and the fundamental principles in ontology design. This is because every ontology methodology ensures these five criteria in the development stage. This has made his criteria a potential evaluation principle for assessing the quality of ontologies. A review of some of the criteria was conducted, and few were selected to choose a methodology for the design of the Citizen Science ontology. The criteria were selected based on their frequent occurrence in literature. Their frequent occurrence makes them usable and acceptable in the semantic web and ontology development processes. The following criteria were obtained from documentation on ontologies and ontology designs the subsections shows a review of the selected criteria and how they are considered in literature for building specific ontologies.

Reuse Capability

Ontology reuse capabilities are the ability of a design methodology to inculcate reusing of existing ontologies in the development process (M Uschold et al., 1998). By delineating classes of particular knowledge in a domain and errands inside these domains, ontologies give structure to understanding which parts of the domain are reusable between different domains. Since building ontologies are tedious and time involving, reusing existing ontologies reduce the overall time and cost involved. As discussed by Ding, Lonsdale, Embley, Hepp, & Xu (2007), there are numerous studies on the reuse and repositories for ontologies. Although these repositories contain rich and well-constructed ontologies, they are for specific domains. However, the domain of citizen science is broad and comprises of many different subdomains. Therefore, ontologies for such domain should be more generic to depict well most of the knowledge in the domain. It is therefore important to reuse existing ontologies as concepts in the geo-citizen science ontology. Reusing existing ontologies in building a new ontology has several advantages. Reuse reduces the engineer's labour involved in enacting the ontologies from scratch (Perez et al., 1999). Since most existing ontologies have already been tested, reusing them improves the quality of the envisage ontology (Ding et al., 2007). However, (M Uschold et al., 1998) also suggested that the existing ontology will share concept/vocabulary with the new ontology. This act will make the mapping between the common component an easy task for the

engineer. Figure 3-3 aims to give a general Overview of the concept of reuse considering three different ontologies. From Figure 3-3, Ontology C can be said to comprise of different ontological concepts/classes from Ontology A and C to describe the intended domain. This is most often than using the same as Relationships.

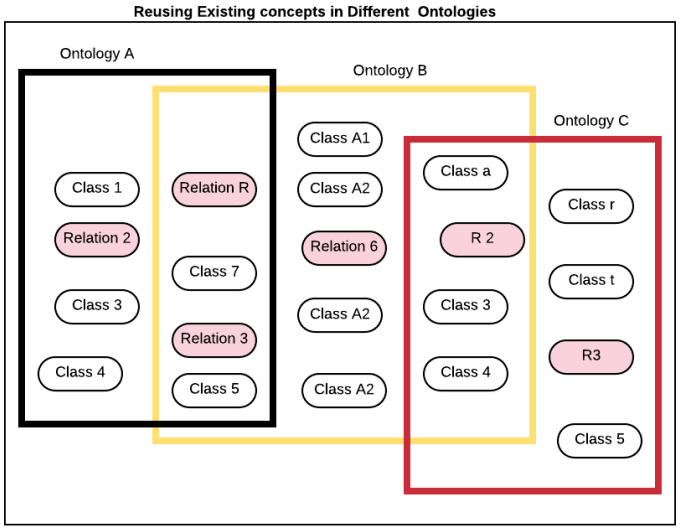

Figure 3-3: Reusing Existing Concepts from Different Ontologies.
Source: Author

Consistency / Commitment

Ontological commitments and consistency are set of coherent concepts that ease an efficient communication about a subject or knowledge. In effect, commitment in an ontology is the ability for a designed ontology to perform the desired action consistently without any contractions with the terms, knowledge and concepts expressed(Newell, 1981). Ontological commitments are mostly perceived as ontological notions. Logically, a typical ontology characterises the vocabulary with which enquiries and statements are traded among specialists. Ontological duties are assertions to utilise the common vocabulary in a sound and steady way. One can conclude, that a pledge to a common ontology is an assurance of consistency. The idea of ontological consistency and instance checking is to see the ontology as static affirmations, which must act naturally reliable, and to which a given occasion state must acclimate (Mike Uschold, 1996a). Most methodologies use some constraint rules as a form implementing consistency constraints and instance checking. These actions are most often referred to as good practice.

Geospatial Capability

Geospatial reasoning in semantic learning plays an important role to the prerequisite for modelling, visualising and envisioning multimodal spatial information, and is exceptional in offering joined examination that incorporates spatial, temporal and topological estimations of information knowledge and data (Brost et al. 2014). Geospatial incorporation in ontology development provides a comprehensive ability to include integrated analysis from multiple forms of spatial information and knowledge for capturing concepts in any domain. Several approaches emphasise the use of semantics to integrate, share, and analyse multimodal geospatial information. These approaches in the ontology design capture inherent spatial concepts and their relationships. It is a potential criterion to consider when designing systems that operate or use spatial information. It gives an expert a suitable structure and potential means of capturing spatial information. Figure 3-4 aims to give a general Overview of all the encoding of geospatial information capability. From

Figure 3-4, concepts which describe spatial information are expressed in a specific geographic coordinate (Latitude & Longitude) to describe or define spatial information explicitly.

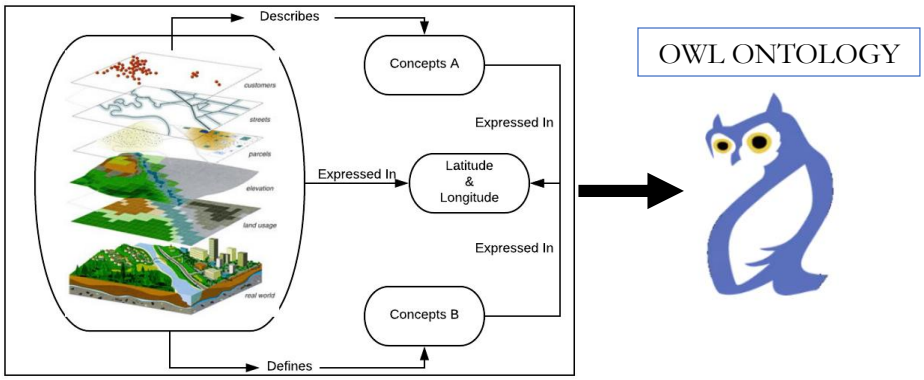

Figure 3-4:Expressing Spatial Information into Concepts in an Ontology. Source: Adapted from (USGS, 2017)

Formularisation

Formularisation in ontology design is the act and process of extracting knowledge from any source for the classes defined in the ontology development process. As discussed by Sintek et al. (2007), the act of extracting knowledge in a domain is the core for the ontology development (Visser et al. 2002). However, the Ontology engineer can determine the type of medium to extract the knowledge from (Poli, 2003). The formularisation process is mostly done in three of four steps depending on the source of knowledge. These steps include defining the source for knowledge extraction, defining sets of operation, selecting a platform which includes tools and language for performing the operation on the extracted information, accepting suggestions, defining mapping rules and finally enacting the rules from the diver domain of Citizen Science. It will be more prudent to consider the diverse sources for the knowledge extraction in other to depict the domain under consideration. Therefore, a sound, methodology that consider diverse modes of formularisation will be appropriate for the Citizen Science ontology. Figure 3-5 aims to give a general overview of the concept of formularisation in ontology design. From Figure 3-5, different concepts are organised and merged into class subclass hierarchy to define an abstract concept in a domain. Concept A is considered to have a subclass of concept B due to the semantics and relations that occurs between the two concepts.

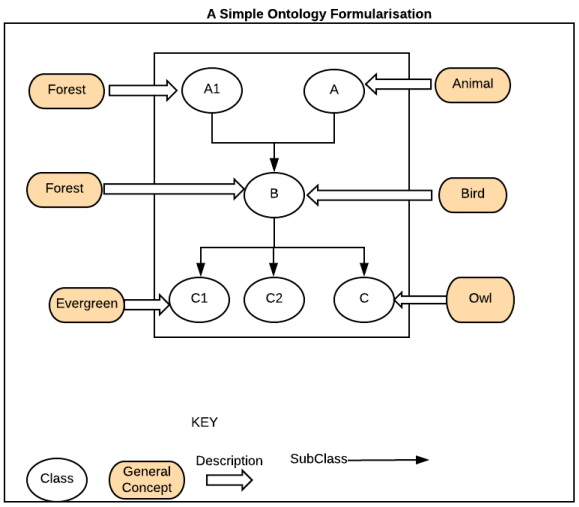

Figure 3-5: The Concept of Formularisation in an Ontology Design. Source Author

Completeness

Occasionally, there are objections to using completely for domain ontologies. This is because of the wide range of knowledge that exists in a domain. We can hardly denote or capture all the knowledge in a specific domain since almost all domains are still under exploration and research. However, a method for the design of an ontology always refers to completeness as a measure of how concise statement in the vocabulary is to determine a specific knowledge**.** Completeness is if questions such as "the extent to which the ontology requirements are met" are used to answer questions related to the domain (Jarrar et al., 2008). What semantic component is needed or missing in the ontology to answer a given competency question clarifies the completeness of the ontology (Mike Uschold, 1996a). An appropriate methodology for the design of the ontology should discuss ways of ensuring semantic completeness in the ontology design process.

Interoperability

Interoperability is the capability of more than one platforms to exchange information efficiently and effectively (Thessen et al., 2011). This ability allows the system to understand the information shared to produce the intended result automatically. Interoperability is an essential characteristic of an ontology design methodology. Some methods for designing ontology support interoperability between systems. This makes ontologies designed with such methodologies efficient and allows reuse since the ontologies share a common foundation and structure (backbone). Because achieving interoperability involves both systems to have a common structure or model for information sharing, ontology methodologies with interoperability capability is a good basis for designing any ontology. This foundation or skeleton allows systems to have common understanding and structure for easy integration with other ontologies that lack such structure (Luciano et al. 2008). This ability allows the ontologies to have easy sharing of knowledge for communication when merging them. An ontology designed with this quality allows interoperability of datasets as well.

. *Figure 3-6* Aims to give the general notion of the concept of interoperability for different datasets. From Figure 3-6, Different information from different sources (Land Use Information, Insects, Birds) that describe different concepts are operated on them making them compatible.

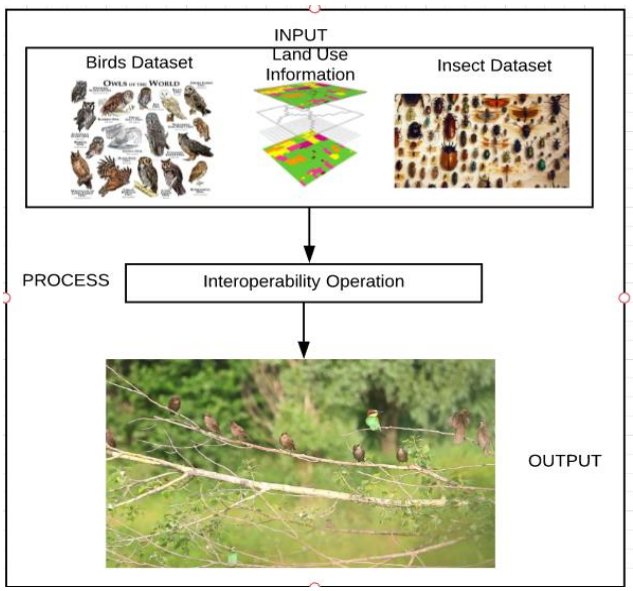

Figure 3-6: The General Notion of making different Datasets Interoperable. Source: Link1[1]/link2[1]/(USGS, 2017)

Co-creation Support

Co-creation in the context of ontology design has been described as the act of collectively creating the ontologies with different people to ensure share responsibilities and combined knowledge from the parties involved. It is mostly used in design processes and popularly termed as co-designed or participatory design, where designers involve the end users and stakeholders in the design process. In ontology design, as discussed by (Sanders et al., 2008), co-creation is the way of ensuring the involvement of the domain expect from start of the ontology life cycle to the end of the final product and most at times maintenance of the ontology (Euzenat et al., 2007). The motivation for this style of design is the ability to render the ontology accepted by the end user and to capture and present the intended purpose of the design (ontology). It inspires domain users and expects to adapt and contribute their ideas to increase the accuracy of the ontology since they are actively involved. Bleumers et al. 2011 argue that co-creation ontology design is crucial and useful since it can provide solutions to the accuracy problems in the envisaged ontology. For a practical and high-quality ontology, a useful criterion for an ontology methodology is to involve co-creation in its development cycle.

Modularisation

Modularization is a non-particular thought that is naturally fathomed as insinuating a condition where at the same time a thing can exist as a whole but can also be seen as a set of parts (the modules) (Grau, 2010). Modularization helps in complexity management, understandability, Context-awareness, and Personalization and Reuse. Versatile quality is an unavoidable part of mainly every design. Stresses over versatility and interoperability of ontologies have delivered significant energy for modularisation from the semantic web gathering (Pérez et al. 2008). Figure 3-7 represent how parts of an ontology can be separated into different ontologies and later been merged into the main ontology.

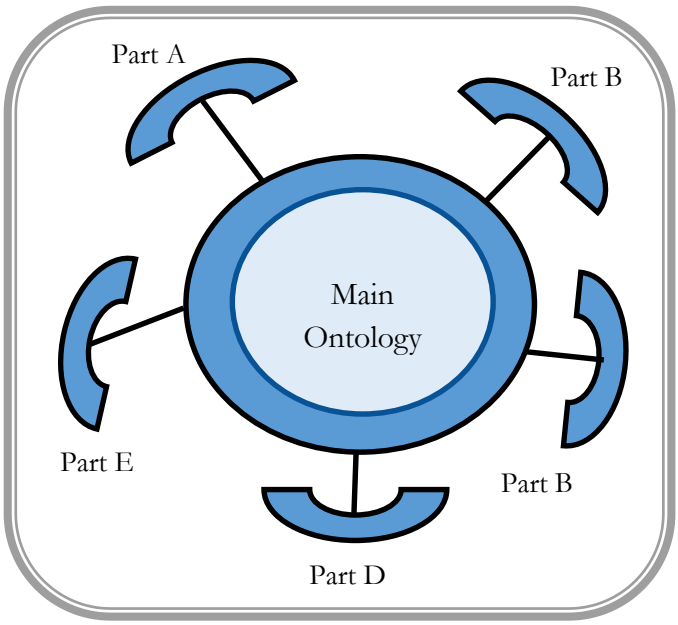

Figure 3-7: The Notion of Ontology Modularisation. Source: Author

3.3. Selecting a Methodology

Several studies are establishing the connection between solving problems of semantic heterogeneity and non-interoperability using ontologies. A general overview by López (1999) states that the efforts to build a

good ontology has resulted in several methodologies for its design. These efforts started back in the 1970s, with the advent of relational database design (Fonseca et al., 2007). Some of these development efforts use existing ontologies (Jarrar et al., 2008) to form a generic ontology. These numerous methods for creating ontologies reported in literature are designed to meet a specific aspect of the ontology characteristics. These methods can be categorised into two broad groups based on their relationship with geospatial issues. The first group is those that consider geospatial component, and others that do not consider geospatial issues. However, it can be argued that the spatial consideration methodologies also consider most of the non-spatial components. Some aspects shared among the two categories are co-creation, reuse and formalisation aspects, just to mention a few. From how diverse the domain of Citizen Science is, a concise methodology that considers both the geospatial and non-geospatial components for creating ontologies will well present and model most of the knowledge in this domain. An analysis of some of the existing methodologies is as follow.

Among the methods used for non-geospatial ontology is the "Developing Ontology-Grounded Methods and Applications" (DOGMA) methodology, which suggests that in creating ontologies; there is the need to consider the influence of usability perceptions on ontology axioms of the domain as well as reusability of the application (Jarrar et al., 2008). Unlike the generic methodology for ontology development, it failed to provide a precise framework to be followed when designing the ontology.

The Methontology methodology is another approach used in the non-spatial ontology design. It proposes a set of guiding principles and criteria to create ontologies from scratch for any given domain. The Methontology methodology identifies the reuse of existing ontology as a key component of efficient ontology design (Fernández-López et al. 1997). The generic methodology for ontology development provides a structured and concise way of generating ontologies from scratch. Unlike the Neon methodology, the generic methodology provides a concise and flexible procedure to be followed in the creation of the ontology. It considers the aspect of both reuses, formalising and co-creation in the ontology development phase. It considers both top-down and both-up approach to specifying concepts in the ontology development. It also supports ontology alignment when merging different ontologies for a specific task. The only diminishing aspect is the consideration of geospatial component in the development phase. The world wide web consortium(W3C) geospatial vocabulary for the design of spatial ontology is among the spatial ontology design methodologies. In this vocabulary, GeoRSS feed is adopted, which is a contribution from many organisations including the open geospatial consortium (OGC). The W3C and the OGC discuss the advantages of extending the W3C geo-vocabulary (Lieberman et al. 2007). Their methodology concentrates on the technicalities of spatial data leaving behind that of non-spatial data. This limitation calls for the need to edit the library by extending it vocabulary for a particular purpose. Based on the selected criteria for the envisaged ontology, Table 3-1 shows how each of the methodologies considers the developed criteria.

The list of reviewed criteria applied to the selected methodologies explicitly shows that no single methodology incorporates all the selected criteria in its design. However, some methodologies involve most of the selected and unselected criteria for their design. Also, the W3c geo and the generic ontology development framework has almost the same number characteristics. However, they have different means of formularization with can have technical implications on the diversity of knowledge can express. The W3C GEO formularization considers more of spatial encoding as compared to the generic development framework. Therefore, the Generic Ontology development framework seems most appropriate for the design of the Citizen Science ontology. The Framework consists of three prominent stages with sub-steps to thoroughly guide the design of a specific domain ontology details can be found in Chapter Five.

Table 3-1 General Overview of Selected Criteria Applied to the Reviewed Methodologies. Source: Author

Method	Reuse	Consis-tency	Geo-spatial	Formali-zation	Com-pleteness	Inter-opera-bility	Co-creation	Modular i-zation	Level of coverage	Sources
Neon	Y	Y	N	Y[7]	Y	N	N[8]	Y	General.[9]	(Oberle, 2014), (De Nicola et al., 2016), (Casellas, 2011)
Methontology	Y	Y	N	Y[10]	Y	N	N	N	Specific[11]	(Jones et al., 1998), (López, 1999)
RapidOWL	N	Y	N	Y[12]	Y	N	N	N	General	(Auer, 2006), (Gil et al., 2014)
OPON Lite	N	Y	N	Y[13]	Y	Y	N	N	General	(De -Nicola & Missikoff, 2016)
Generic ontology development framework	Y	Y	N	Y[14]	Y	Y	Y	Y	General	(Rajpathak et al., 2011),
Normalized Methodology	Y	Y	N	Y[15]	Y	Y	N	N	General	(Buccella et al., 2011)
Onto-Agent	N	Y	N	Y [16]	Y	N	N	N	General	(Hadzic et al. 2009)
HCOME	Y	Y	N	Y	Y	N	Y	Y	General	(Kotis et al., 2006)
SMOL	Y	Y	N	Y	Y	N	Y	Y	General	(Gil et al., 2014)

[7] A formal or semi-computable model was used
[8] No discussions were made on issues regarding this quality indicator
[9] The methodology develops ontology base on scenarios
[10] No Specific formalization tool was recommended
[11] Developed for Knowledge-based systems
[12] All knowledge is based on RDF triples
[13] UPON Lite was recommended as a tool for formalisation
[14] All knowledge is based on RDF triples
[15] All knowledge is based on RDF triples
[16] (First-order logic)

TOVE	Y	Y	N	Y	Y	Y	Y	Y	General	(Casellas, 2011), (Perez et al., 1999)
W3C Geo 2003	Y	Y	Y	Y[17]	Y	N	Y	Y	General	(Lieberman et al., 2007), (W3C, 2007)
Unified Methodology	Y	Y	N	Y[18]	Y	Y	Y	Y	General	(Mike Uschold, 1996b)

*The dark colour with Y represents Yes which means the Methodology considers the proposed criteria
*The lighter colour N represents No which means the Methodology considers the proposed criteria

[17] GeoRSS Feeds in the form of triples.
[18] All knowledge is based on RDF triples

3.1. Related and Relevant Efforts in Designing Ontologies

This section is a review of relevant works on ontology design that captures concepts in the domain of Citizen Science.

There are different ontologies designed and implemented according to literature. Most of these developed ontologies are intended to enhance the exchange of information across various platforms and to make information machine-readable (Kotis et al. 2006). Most of the existing ontologies are intended for different domains and applications. Examples of the domains having ontologies to express their knowledge are the business industries, biological, healthcare, government and telecommunication agencies. There is no precise designed ontology to represent different dataset in the domain of Citizen Science. However, several different ontologies express specific concepts in the domain of Citizen Science. This section looks at some of the existing ontologies that can be considered to express specific knowledge in the domain of Citizen Science. These ontologies are grouped into two based on the use of spatial information in their design process. These two groups are the non-spatial ontologies and spatial ontologies. The list of ontologies considered is based on their relevance to the domain of Citizen Science and their frequent usage.

3.1.1. Non-Spatial Ontologies

British Broadcasting Cooperation (BBC) has several ontologies designed to express different sections of their work. Some examples include the Business News Ontology, Creative Work Ontology, Journalism Ontology, Wildlife Ontology and others. Among these ontologies, the Wildlife Ontology (WO) stands out to be a precise formulation of vocabularies that captures biological information which can be considered under Citizen Science (BBC, 2017). The wildlife ontology is designed to express knowledge on biological species. It aids in publishing data of all forms of biological species and their related taxa on the web by describing the relation of species to their natural environment (habitats). However, the ontology does not describe individual species but defines the notion of a species based on their characteristics and their relations to the environments. The BBC Wildlife Ontology provides a general understanding for grouping biological species according to their characteristics and their environment. Unlike the BBC wildlife ontology, the Biological Taxonomy Vocabulary (BTV) considers species as classes but not the notion of species in their classification system (Schulz et al. 2008). The BTV describes the domain based on the identification of a group of organisms that relate to a specific study area. Another ontology for the biological domain is the GeoSpecies ontology (NCBO, 2012). It consists of vocabularies designed to describe some of the biological species like in the case of both the BBC Wildlife ontology and the biological taxonomy vocabulary. This ontology depicts a universal classification system for biological species. Each species is organised into their respective kingdoms, phylum, class, order and genus. Unlike the BBC ontology, the Geospecies ontology describes only the taxa of biological species. The Vertebrate Taxonomy Ontology (VTO) is a hierarchical taxonomy of vertebrates' organisms designed for the integration of biodiversity data semantically (NCBO, 2017). It describes details on the information of both extinct and extant hierarchy of vertebrates with both their common names and scientific names. The VTO helps in understanding the phylogenetic relationships that exist among vertebrates. The hierarchy for vertebrates serves as the backbone for the National Centre for Biotechnology information in the United States (NCBI) taxonomy. The plant ontology consortium developed the Plant Ontology (PO), and it is a collection of ontologies that describe the domain of plant anatomy, plant growth, and plant development (NCBO, 2017). The plant ontology was designed for education and semantic applications that require knowledge in the domain of floral. It captures knowledge that ranges from plant roots system to plant structure and plant anatomy. Most of these ontologies describe the species in relation to their environment. Moreover, there are well-developed ontologies designed explicitly for environmental characteristics. An example is the Weather ontology (Staroch, 2013). The

weather ontology describes both lower-level, middle-level and upper-level domain-specific concepts in the Wether forecast domain. The lover-level formalise concepts such as different measurement systems in the domain of weather forecasting. The upper-level concepts express phenomenon generalised to be classified as a subdomain such as weather phenomenon. The middle-level concepts express general concepts that are defined in the upper-level concept formularisation. Examples are rainfall considered as a weather phenomenon. All the above ontologies can be considered relevant and useful for the design of the envisaged Citizen Science ontology. Therefore, Classes and properties from these ontologies will be considered and used to emphases the reuse component of the selected methodology.

3.1.2. Spatial Ontologies

Ontologies for expressing elements of different knowledge artefacts in the domain of geospatial information using spatial relations have become an appropriate means of sharing spatial data on the semantic web (Stuckenschmidt et al., 1999). Ontologies developed on the bases of spatial relations and spatial objects are typically referred to as geospatial ontologies. Examples of such ontologies include the Ordnance Survey Ontologies, W3C Geospatial Ontologies and Geographic entity ontology. This section describes efforts made towards the development of geospatial ontology and encoding of geospatial information on the semantic web. The notion of geospatial ontologies is to model spatial entities as objects and use spatial relations to map these objects in the form of triples. Geospatial ontologies such as the ordinance survey ontologies are expressed in a formal language to capture knowledge such as the geometry of spatial objects, and postcodes. As shown in Figure 3-4, different spatial concepts can be encoded and mapped to concepts in a domain. Ontologies can map and groups spatial entities as objects based on their geometry. Table 0-2 in the Appendix gives detail discussion of different efforts and description of some existing spatial ontologies and an overview of the encoding and spatial relations used.

In general, the methods to engineer an ontology can be summarised as top-down: from generalisation to specification, or, bottom-up: from specification to generalisation, or, middle-out: from the essential concepts to generalisation and specialisation (López, 1999)). The design of the Citizen Science ontology will consider the middle-out approach, where the most important concepts will be expressed, and both the general and specialised concepts will be obtained. The most specialised concepts will be adapted from different existing ontologies emphasising the reuse component of the selected methodology (Generic Ontology Development Framework). The Ontology will focus on spatial relations described by the Ordnance Survey spatial relation ontologies and W3C Geospatial encodings for geometric features.

4. FRAMEWORK AND USE CASE

This section discusses the selected methodology from chapter three and other relevant frameworks considered in the design of the ontology. It discusses the use of spatial relations for designing ontologies and finally concludes on some selected uses cases to serve as proof of concepts to verify the relevance of this project.

4.1. Framework

The design of the ontology is based on existing frameworks and algorithms to ensure a sound flow and coherent design. This section aims at given a general overview of the list of frameworks adapted for the design of the ontology. It concludes with a description of the different formal languages used for expressing and querying ontologies.

4.1.1. IEEE Standard for Software Development Life Cycle

There are several frameworks for designing coherent software applications. Examples of these frameworks include the IEEE Standard for Software Development Life Cycle (IEEE, 1991) and the Agile Framework (Auer, 2006). Among these frameworks, the IEEE Standard for Software Development Life Cycle is a potential framework that provides clear and precise structure for building software applications. This framework offers set of activities that establish the underline processes required for the development and maintenance of software (IEEE, 1991). It provides a set of activities for examining the purpose of the envisaged software application. The process of the examination leads to a coherent understanding of the application under consideration (Citizen Science ontology). The framework provides a thorough software design process by which the software requirements are directly translated to the into representations and sections of the envisaged
software components.
The framework has several development stages that when combined constituted the overall flow of the Citizen Science ontology. Figure 5-1 gives a general overview of all the component of the Framework. However, the framework is a complex multiphase. Therefore not all the different component will be considered in the design process of the ontology. The stages considered includes the following

1. **Management Activities Section;** This section discusses the overall management activities of the designed ontologies. These activities will include identifying the required resource for modelling and defining steps for creating the ontology. The developed criteria from this section serve as the foundation for creating the ontology. The management includes a grouping of ontology classes based on a higher-level abstraction to accommodate the different part of Citizen Science domain.
2. **Development activities section;** This section considers the process of the actual design. Due to the flexibility of the IEEE Framework, the whole of this section is replaced with the selected methodology in chapter three (Generic Ontology Development Framework). The generic ontology development framework as the selected tool has several design components. These components are fully discussed in chapter five.
3. **Supporting Activities Section:** This section gives an overview of the implementation strategy and a general overview of the quality testing for the designed Citizen Science ontology.

However, the Generic Ontology Development Framework is incorporated with different formal languages for expressing the knowledge in any domain of interest. The next section aims at given a general overview of the different formal languages used in ontology design.

4.1.2. Formal Ontology Languages

There are specific formal modelling languages for expressing ontologies. These formal languages are modelled to ease sharing and reusing of information across different systems/platforms. Example of these formal languages are the Defence Advanced Research Projects Agency *(*DARPA) Agent Markup Language, Knowledge Interchange Format (KIF), The Resource Development Framework (RDF) and Resource Development Framework Schema (RDFS) combined in the ontology language layer (OLL). This section describes OWL and RDF and RDFS as a formal language with much expressiveness for easing the sharing and reusing of information across different platforms.

OWL is formal rich, expressive language designed by the W3C for ontologies. Ontologies expressed in OWL are distinguished by formal semantics and RDF / XML- based serialisations for the Semantic Web. OWL contains sublanguages such as OWL DL and OWL Lite which are syntax with high-level abstraction. They are mostly implemented in ontology and semantic editors such as Protégé with many standard reasoners which include Pellet, RacerPro, FaCT and HermiT. Examples of ontologies designed in OWL include The Friend Of A Friend (FOAF), Upper Mapping and Binding Exchange Layer (UMBEL) and the Dublin Core Ontology. The **FOAF** ontology is a vocabulary which describes people and objects in social networks; the ontology expresses the relationships that exist among persons, objects and both in the social network. The **UMBEL** ontology is a vocabulary for the advancement of ontologies being planned for interoperation and gives a reference structure of various thoughts that give a system to meet and interoperate datasets. **Dublin Core** ontology is a vocabulary for describing metadata of web documents, physical resources and other objects.
Another standard for modelling knowledge representations is the Resource Description Framework (RDF), which is a recommendation by the W3C. In RDF technology, the term resources are mostly used to represent instances (an example or single occurrence of thing). Resources may have a subjective number of properties (attributes with literals, e.g. numerical values, strings) to the resource or relations linking two resources. In most Knowledge representations, a statement termed triples are used to represent resources and their attribute. Triple consist of three fragments namely subject, predicate and object.
These three components help in identifying the relation and properties that exist among resources. In RDF technology, the term Unified Resource Identifiers (URI) is used to identify resources and their properties. However, URI has substrings at the beginning to avoid frequent relapses of the same strings. These substrings are replaced by prefix named namespace setting distinction among resources. These distinctions are set across all entities (concepts, properties, and individuals).
RDF is built up by RDF Schema (RDFS), a recommendation by the W3C for adding features that go beyond only the expressive power of RDF. The Schema provides a set of concepts and properties with prefix RDF. Table 4-1 gives a review of some concepts and properties with RDFS as prefix and their corresponding description proposed by the W3C. Table: RDFS Prefix for describing resources and properties as well as RDF classes.

Table 4-1: Examples of Property Names with Domain and Range. Source: (W3C, 2014)

Property name	comment	domain	range
rdf:type	The subject is an instance of a class.	rdfs: Resource	rdfs: Class
rdfs:subClassOf	The subject is a subclass of a class.	rdfs: Class	rdfs:Class
rdfs:subPropertyOf	The subject is a subproperty of property.	rdf: Property	rdf: Property
rdfs: domain	A domain of the subject property.	rdf: Property	rdfs: Class
rdfs: range	A range of the subject property.	rdf: Property	rdfs: Class

Table 4-1 gives a general idea of some of the properties with RDF prefix, the domain and range for the said property name. A property name comments refer to the domain rdf:Resources and ranges over most rdf:Literals. A general comparison of RDFS and OWL reveals that they both share many concepts and properties. However, OWL is based on RDF and RDFS and has more expressiveness for modelling

knowledge on the semantic web. Table 4-1 shows properties and concepts with owl prefix for modelling ontologies in OWL as proposed by the W3C. A detailed description of the owl syntax can be found at (link). The languages used to model ontologies and knowledge representation are implemented in frameworks which have adequate criteria for modelling ontologies. RDF formats are structured to store information and data in a directed label graph on the semantic web. These graphs are mostly invoked or queried to realise resources and determine relevant information in these directed graphs. These queries are performed by an expressive language called SPARQL.

SPARQL stands for SPARQL Protocol and RDF Query Language. It is a query language designed by the W3C for retrieving information from RDF data structure. It is a semantic query language designed with the RDF and RDFS syntax for easy manipulation of RDF data to retrieve potential information from diverse data sources. Just like Structured Query Language (SQL) operating on relational databases to retrieve valuable information, SPARQL operates on graph data (RDF) from different sources. These data sources are in the form of triples (Subject Predicate Objects). SPARQL performs a different form of operations such as aggregations, conjunction, disjunctions and other forms of data manipulations techniques that enable extraction of relevant information from graph data. Considering Figure 3-1, a SPARQL query can be used to operate on the model to retrieve information such as all instances of OWL species that has habitat Swampy Coniferous Forest. The SPARQL query will select the graph portions that describe the relation of ***HasHabitat***. It will then make logical conjunction of all Owl individuals that are within the selected graph pattern. The result of SPARQL queries are data frames of sets or another RDF graph describing the required information. In effect, SPARQL enables queries over data with "Key-Value" (JSON). Figure 4-1 shows a simple query language constructed to retrieve information on different species in an RDF data. From Figure 4-1, The select clause specifies the items and values from the SPARQL endpoints. The where clause clarifies the condition for the selection. These conditions determine the different triple patterns that must be selected (i.e. All selected endpoints are values that meet the where condition). The capability of SPARQL includes its ability to process both spatial and non-spatial relations. However, the geospatial querying component of SPARQL is termed GeoSPARQL. The next section gives a general overview of spatial relations that was considered in the design of the ontology. Spatial relations were considered as predicates both spatially and as attributes in the design.

```
PREFIX CS:<http://www.semanticweb.org/yawfrimpong/
          \ontologies/untitled-ontology-13#>

SELECT (Distinct ?o as ?Land_Information)
WHERE
{
?s ?p "San Francisco" .
?s1 CS:FoundOn ?o .
?s1 CS:HasHabit ?Forest .
}
```

Figure 4-1: Simple SPARQL Query. Source: Author

4.2. Spatial Relations

An essential capability of geographic information systems that distinguishes it from other information systems is its ability to process spatial relations among spatial entities. Spatial relations provide means of describing features and entities pertaining to a location. Examples are the distance between the capital of Ghana and the capital of Nigeria. The distance separating residential buildings from reserved farmlands at Amasaman in Ghana. These relationships indicate the distance between the two entities. This section seeks to give a general overview of some spatial relations that will be considered in the ontology design.

Spatial relations often serve as a predicate that links two spatial entities. The results when checked returns an either true or false (Boolean). However, not all spatial relations are considered to yield Boolean results. Some produce geometries that express the relationship among the given geometries. Most of these predicate compares the point set (coordinates) of the two geometries to check for the relation concerning their location. We can compare two geometries using spatial relations such contains; the contains relation can be confirmed by comparing the point set of the two geometries. If the point set of the first geometry is entirely part of the point set of the second geometry, we can confirm that the second geometry contains the first geometry. An appropriate means of checking the spatial relations is by comparing the interiors, exteriors and the boundaries of the two geometries using the nine-dimensionality intersection model developed by Egenhofer and Herring (OGC, 1999). The nine-dimensionality intersection model compares the geometries using a pair-wise mathematical model underlining the model. Table 4-2 shows the nine-intersection models and how they are applying to two given geometries.

Table 4-2: The Nine-Intersectional Model. Source: (Egenhofer et al., 1991)

		Geometry B		
		Interior (I)	Boundary (b)	Exterior (e)
Geometry A	Interior (I)	dim(I(**A**) ∩ I(**B**))	dim(I(**A**) ∩ b(**B**))	dim(I(**A**) ∩ e(**B**))
	Boundary (b)	dim(b(**A**) ∩ I(**B**))	dim(b(**A**) ∩ b(**B**))	dim(b(**A**) ∩ e(**B**))
	Exterior (e)	dim(e(**A**) ∩ I(**B**))	dim(e(**A**) ∩ b(**B**))	dim(e(**A**) ∩ e(**B**))

4.3. Use Cases

This section considers problems in Citizen Science data integration that needs the attention of the envisaged Citizen Science ontology. It describes the problems and how the ontology will be used to provide a solution or an intermediate solution to the problem. The use case ranges from almost all aspects of Citizen Science under consideration in this project. Citizen Science ontology. It describes the problem and how the ontology will be used to provide a solution to the problems.

Use Case: Inculcating Biodiversity Conservation Planning into City Conservation Planning

Biodiversity is the diversity, of animals and plants, the number or abundance of different species and other living life forms in a specific territory (Willig et al. 2017). Biodiversity conservation aims at given the spatial distribution of the different species at a particular time. Reports from biodiversity conversations state that various species are often conserved and protected from extinction due to their inability to cope the fast-changing climate (Margules et al. 2002). Assessing potential impacts of biodiversity conservation in city planning policies are means of protecting the environment as well as keeping some vulnerable species from extinction. Therefore, conservational city planning utilised spatial locations of species during planning for consistent quality of living space (Hanski, 2016). The involvement of biodiversity in city planning is an indication of how well a city is planned(Graham et al. 2015). There are many useful tool and mechanisms for joining biodiversity conservation into conservational city planning (Hyder et al., 2015) Example of such

tool is the stochastic patch occupancy models (Hanski, 2016). For such tools to work efficiently during city planning framework, other indicators of species occurrence need to be studied (Graham et al. 2015). However, bringing together different species and species characteristics always result in inconsistencies due to the different semantics and syntax of the datasets. As discussed by Flowerdew, 1991, these incompatibility is most often due to the separate data columns and data fields that exist among the variously selected datasets, rendering them non-interoperable. Therefore, there is the need to solve the non-interoperability issue among the generated Citizen Science. Resolving the non-interoperability matters can help answer questions such as the ones in Table 4-3.

In other to solve the non-interoperability issue using competency questions raised Table 4-3, a standard for integrating the semantics of the dataset is required. This integrated dataset can produce potential information which can be incorporated into city conservational planning policies. Therefore, a medium for solving the non-interoperability in Citizen Science with the Citizen Science ontology will be a means of realising a potential solution to this problem. This use case considers the United States and its regions as the study area for biodiversity conservation incorporation into city planning. The next steps in this section show how different dataset can be made compatible by considering specific classes and relations in the ontology. Three steps are provided to show how the ontology can be used at instance level of the different classes.

Set 1: Validating GeoWiki landcover Inputs in California

Overview: Confirming Geowiki landcover inputs in the city of California (validations on GeoWiki websites using Owl sightings, insect sightings (Dragonflies)). Geowiki is a platform that aims at building a global land cover and land use information for public use by capturing different sceneries and sharing these captured landforms with others. It provides means for the general public to submits different land use and land cover classes to be validated and accepted by different researchers. However, some of the information comes from untrusted and unreliable sources. Therefore, a standard system that can check the submitted inputs from the public will be a potential means of validating the citizen's input on this platform. This use case is to prepared potential information that can be used to validate land cover classes in the State of California using the information contained in the envisaged Citizen Science ontology.

Datasets

1. Owl Sightings: This dataset is in two folds: A shapefile and a CSV file. The shapefile contains information on the location of spotted owls whiles the CSV file provides knowledge of the types and characteristics of the environment of the sightings with possible location description.
2. Dragonflies sightings: The dataset contains the list of spotted dragonflies and their nymphs and the environmental characteristics:
3. Land Use and land cover for California: This is a user-defined land use and land cover datasets from GeoWiki.

Review: In the city conservational planning, a land area with such land characteristics can be assigned with unique laws to protect and conserve such species. The information from land validations and animals that exist at such locations can be vital when delineating land for state ponds and other agricultural projects. Figure 4-2 shows a preview of the different datasets to show their spatial coverage and spatial overlap.

Set 2: Assessing Habitat Condition of Wetland Birds

Overview: There are quite numbers of birds that live most of their life cycle in waterlog areas; these birds usually are referred to as wetland birds. Waterlog areas such as swamps, floodplains and lagoons serve as their habitat and provide food for such species. However, the poor conditions and environmental problems posse a threat to these species in their habitats. Knowledge about waterfowl's food and feeding pattern and

behaviours are fundamental to effective management of waterfowl populations. This use case aims at presenting information for decisions on wetland birds such as the waterfowls by presenting facts about the locations of these species reported under Citizen Science projects and their environmental conditions.

Datasets

Waterfowls sightings: The dataset presents the locations of some sighted waterfowls species. It was obtained from the USGS BISON platform.

Land cover datasets: This dataset presents the land use characteristics of areas in California. The purpose of the dataset is to check the type of land cover for the sighted waterfowls.

Insects datasets: The insect dataset presents locations and the characteristics of insect in wetlands.

A-Z aminal information: A general characteristics and information on different species of animals to their environment.

Review: In city conservational planning, the presence of different species can contribute to the allocation of that area for specific land use. Therefore, if there are huge reported species of wetland bird in a particular area, possible investigations can be conducted, and allocation of such area can be delineated for such purpose. The purpose of the ontology is to integrate the datasets to present information on the environmental and land characteristics of these areas based on the user-generated datasets. Figure 4-3 shows a preview of the different datasets to show their spatial coverage and spatial overlap.

Set 3: Understanding New Ways of Developing Vertical Forest[19] (Biodiversity)

Overview: The act of using vertical forests to nurture biodiversity can be said to be an effective means of preserving both land and unwanted species from extinction in the natural environment. It can be an efficient means of replenishing the natural air in our locality. Therefore, inculcating vertical forest (Vertical Biodiversity) into city conservational planning policies can serve as a means of preserving friendly species that are almost in extinction. This use case will require other additional information before a decision can be reached information.

Datasets

1. Birds sightings from the Avian Knowledge Network. It consists of locations of sighted different species of birds.
2. Land use data from GeoWiki: The land use and land cover datasets are to give the land characteristics of these areas that can be developed into the vertical biodiversity.
3. Existing geodata on city characteristics and landforms
4. Species information from the A-Z animal network. This dataset has no geographic location but contains information on the relationships that exist among different species. Examples of relationships are Diet, Host and parasite.

Review: In biodiversity conservation, different and lost species can be kept in such environment during the city planning. Since the selected recreational buildings are closed to vegetation, there can be a quick and easy adaptation of such species to their new environments. Figure 4-4 shows a preview of the different datasets to show their spatial coverage and spatial overlap. The Reserved areas on the map are location formally reserved which cannot be used in determining possible areas for vertical forests.

Table 4-3 shows a list of competency questions selected to test the quality of the ontology. Each set of review is presented with different sets of the questions for the data integration process.

[19] Vertical Forest is a model proposed for creating sustainable buildings. This to promote reforestation and to contribute to regenerating green environment.

Table 4-3: List of Competency Questions to Test the Quality of the Ontology. Source: Author

List of Competency Questions to Test the Quality of The Ontology	
Number	Question (label is provided as [Q Set. Question no])
SET 1	
Q 1.1	Which region is the highest reported number of species?
Q 1.2	What are the land use and land cover characteristics of those regions?
Q 1.3	Are there any risks of natural disaster reported in these areas?
Q 1.4	Where are the locations of Owl sightings?
Q 1.5	Are the sightings reported Forest cover as proposed by the user input?
SET 2	
Q 2.1	Where are most of the recently reported number of Waterfowl Birds?
Q 2.2	What are the characteristics of such areas?
Q 2.3	What other species are available at those locations
SET 3	
Q 3.1	Which areas can support vertical forest?

Figure 4-2, 4-3 and 4-4 give the impression of the overlap among the dataset for each set of use case. The different colours of points in the Maps show the different species presented in the dataset. From Figure 4-2, the Labels **A**, **B** and **C** represent the different zoomed version of the dataset. **A**- shows the distribution of Species in the United States with emphases in California as shown in **B**. **C** shows a zoomed version.

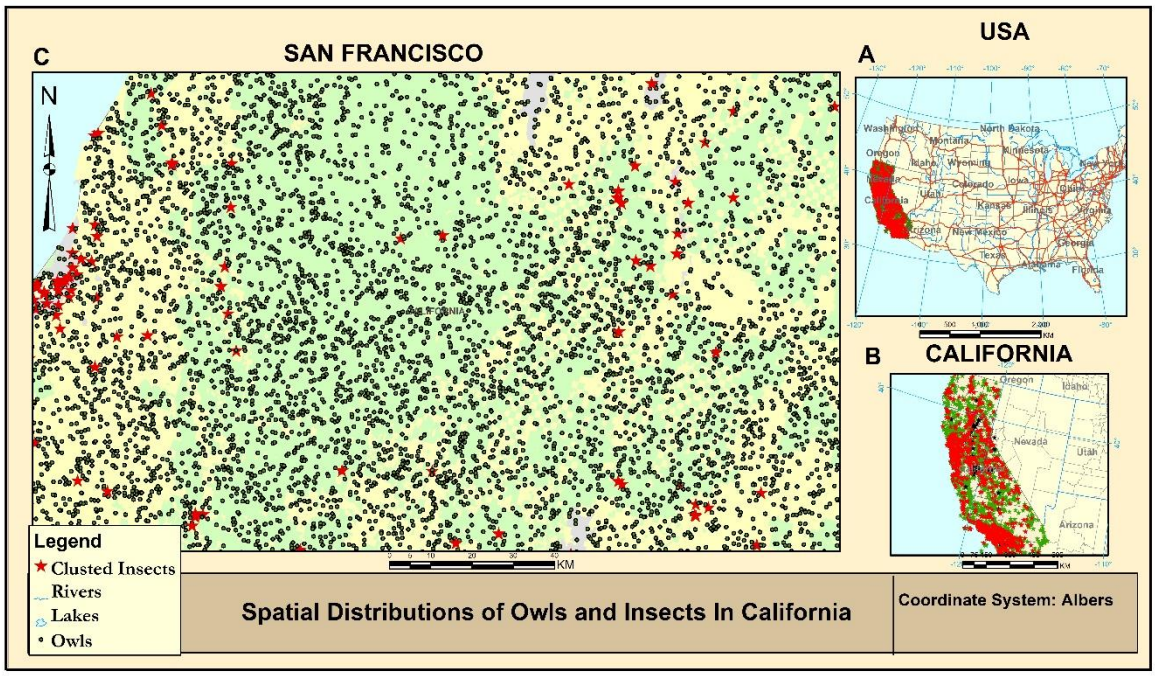

Figure 4-2: Land Cover Validations using Owl(Green) and Insects(Red). Different Colours of Polygons on the Map shows the Different Land Classes that needs Validations. Source:Author

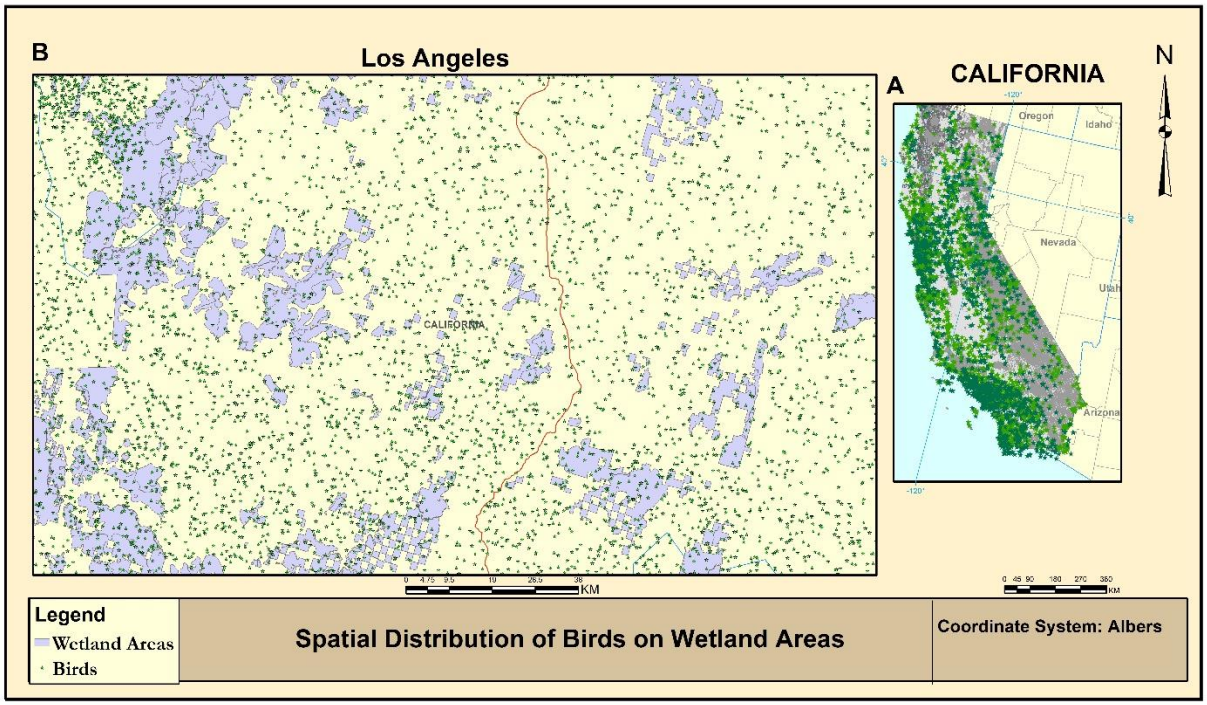

Figure 4-3: A map showing the Distribution of Birds in California for Assessing of Habitat Conditions of Wetlands Birds(Green). Source: Author

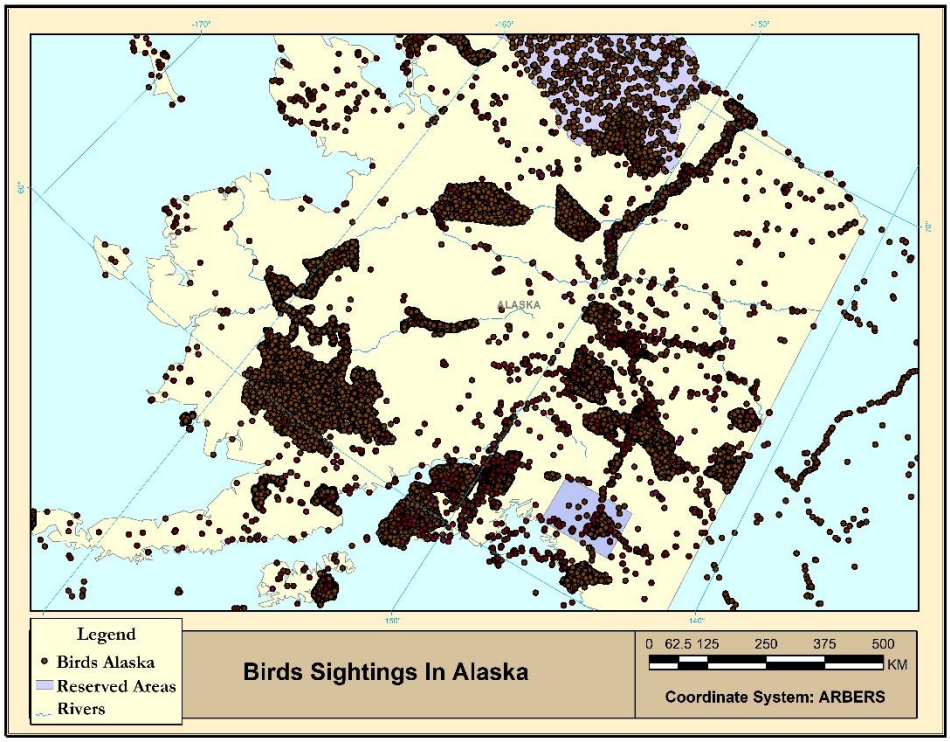

Figure 4-4: Spatial Distribution of Bird Species in Alaska for Developing Vertical Forest. Source: Author

5. DESIGN AND IMPLEMENTATION OF THE ONTOLOGY

This chapter describes the design and implementation of the Citizen Science ontology**.** It discusses the granularity, formularisation and other specific components of the ontology. It finally concludes with a set of implementation strategies for realising its capabilities. The design of the ontology is based on the IEEE standard for software development fused with the generic ontology development framework.

5.1. Introduction

The process of building the Citizen Science ontology is based on the IEEE software development life cycle (IEEE, 1991). The framework is a standard that provides clear and precise structure for building software applications as discussed in chapter four. The overall steps to be followed is shown in Figure 5-1. Three relevant sections are adapted from this IEEE framework. These sections include the Management Activities, Development Activities and Support Activities. Some characteristics of the IEEE framework for software development framework include a well-structured and logical grouping of components that ensure a logical flow of each section in the life cycle process. The framework allows flexibility and consistency; therefore, the selected methodology (Generic Ontology Development Framework) is fused in this framework. Figure 5-1 shows the general overview of the design of the ontology using the two fused frameworks.

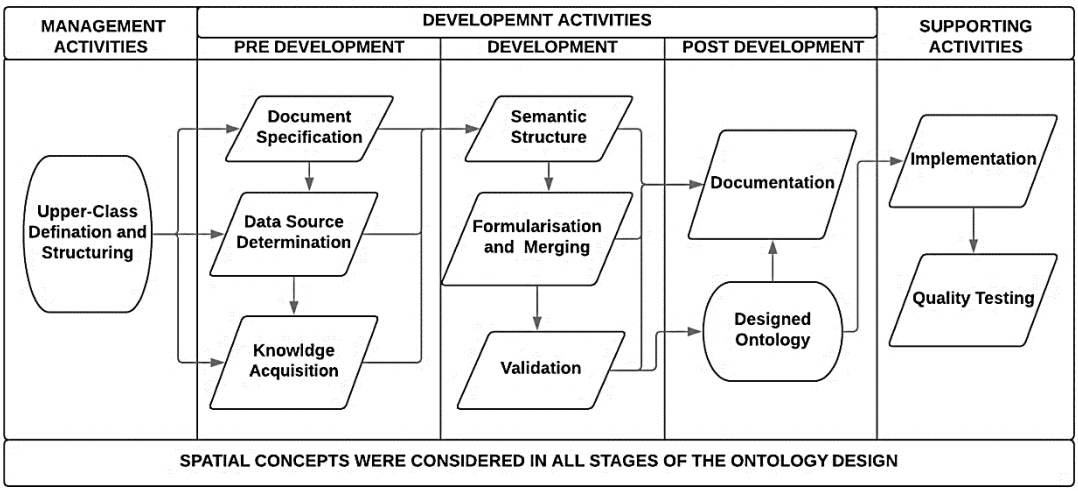

Figure 5-1: Overall Structure of the Ontology Design and Implementation.
Source: Adapted from (Rajpathak et al., 2011)

This section describes the general management activities for the design of the ontology. It aims at organising the ontology according to the activities describing Citizen Science. The management includes a grouping of ontology classes based on a higher-level abstraction to accommodate the different part of Citizen Science domain. (Rajpathak et al., 2011)

The upper level of classes in the ontology serves as a level of abstraction which gives accurate categorisation to the ontology to accommodate different sections in the ontology design for different purposes. However, these levels of abstractions are not referred to any identifiable, concrete entity in the domain of Citizen Science. However, these level of conceptualisations tries to model Citizen Science based on the characteristics and results obtained from Citizen Science projects. The Upper-Level ontological classes selected for the Citizen Science ontology is grouped into seven (7) distinct components. These components

provide means of specifying Higher-Level conceptualisations in the ontology. Figure 5-2 shows the seven higher level ontological classes considered in the Citizen Science ontology. These seven upper classes are **Data**, **Knowledge**, **Projects**, **People**, **Policies**, **Tools** and **Publications**. The upper-class **Data** tries to model different types of data and provides links to some available datasets using the linked data principles. The Upper-class **Knowledge** provides an abstraction for knowledge captured in the domain of Citizen Science. It captures information from different Citizen Science datasets, projects tools publication and many others. The upper-class **People** models the roles and functions of people per project. These roles and functions for different people give an overview of the datasets generated regarding quality. Upper-class **Policy** expresses available policies that govern Citizen Science projects and activities. The **Project** Upper-class expresses the different types of projects that yield different datasets and different knowledge that can be joined to solve a practical use case. The **Publication** upper-class serves as the list of available literature that promotes and describe Citizen Science activities and knowledge. Finally, the upper-class **Tools** serve as the list of tools and technologies used for capturing data in Citizen Science. The higher-level classes are just an abstraction. The design of the ontology at this stage will consider most concept in Upper-Class **Data** and Upper-Class Knowledge. Moreover, spatial components and spatial relations are recursive across all Upper-Classes.

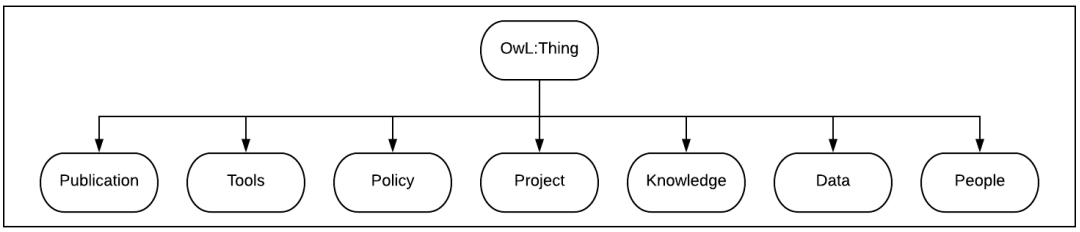

Figure 5-2: Upper-Level Ontological Classes: Source: Author

5.3. Development (Generic Ontology Development Framework)

The ontology development section reports on the design of the Citizen Science ontology using the Generic Ontology Development Framework at the development section for the adapted IEEE framework. Figure 5-3 shows the sections considered at the development stage. All sections are interrelated. The output of the current section forms the Basics of the preceding section. Each tag is defined in the relevant section (A, B and C)

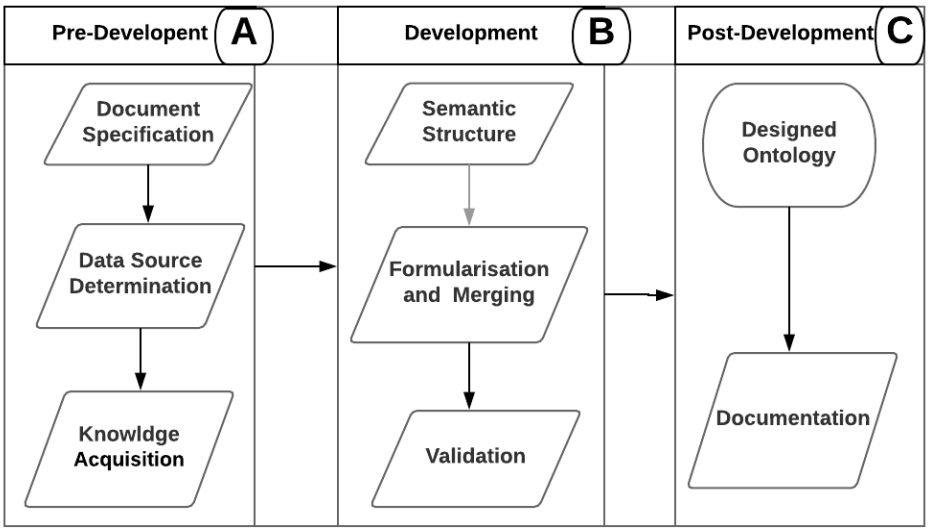

Figure 5-3: General Overview of the Generic Ontology Development Process. Source: Adapted from (Rajpathak et al., 2011)

5.3.1. Pre-Development (A)

The domain of Citizen Science is a broad category that encompasses almost all aspect of both Natural and Social science. To provide a proof of concepts in this project, the Citizen Science ontology considered most of the domain of biodiversity and natural hazards. However, provisions were made for all aspects of Citizen Science to be captured in the designed ontology. Figure 5-4 shows steps followed to acquire all the needed information in the development stage. Form Figure 5-4, the scope gives a general depiction of the broad domain of Citizen Science. It aims at defining an intelligible scope based on the available datasets and information. The purpose of this ontology is to help provide a solution to the problems of non-interoperability in Citizen Science community for reuse of heterogeneous datasets.

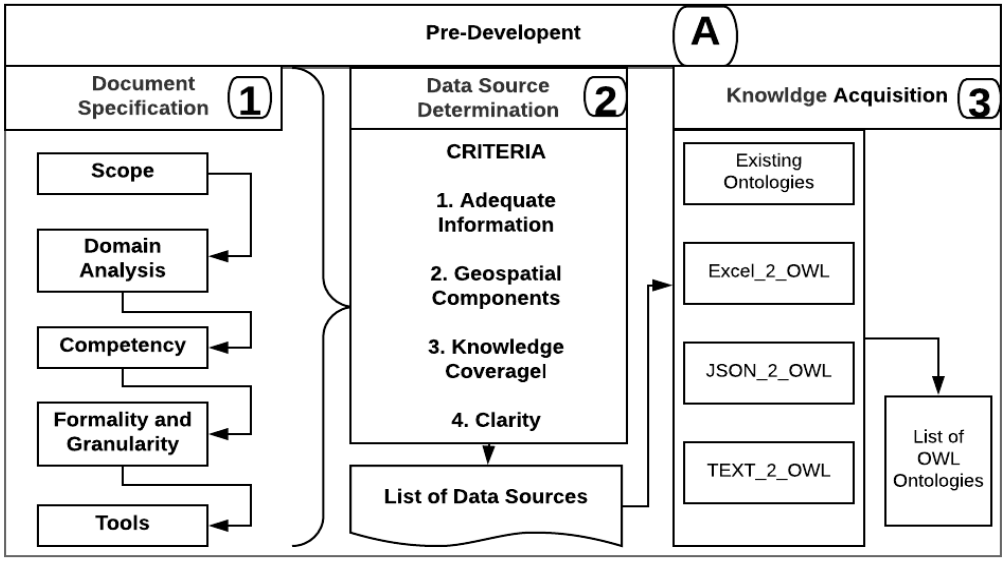

Figure 5-4: Overall Steps at the Pre-Development Section.
Source: Adapted from (Rajpathak et al., 2011)

Step 1 Document Specification

This section describes a document specification of the ontology. The specification is to help realise the intended purpose, scope, competency, granularity and the formality of the ontology. The general analysis of the domain of Citizen Science to select a piece of information serving as a concept for the design of the ontology is highly based on the document specification section. It indicates platforms and algorithms to use in the design of the envisaged ontology. The following were considered in the document specification section:

Scope: Scope means merely the extent to which the Citizen Science ontology captures the knowledge in this domain. In this project, the design of the ontology is to ease data sharing and attempt to solve the problem of non-interoperability in Citizen Science, more precisely non-interoperability among Citizen Science datasets. The scope comprises a wide range of Citizen Science projects. The range of Citizen Science cuts across almost all platforms in science. As a proof of concept, most emphases of the envisaged ontology will consider the domain of environmental and biodiversity. Table 5-1 shows a list of some of the selected projects to be considered with the link to their resource and platforms.

Domain analysis: There are numerous ongoing projects in Citizen Science currently, these projects serve as the basis for designing the Citizen Science ontology. The act of selecting a specific project to be considered in the design of the ontology followed the following criteria. The criteria helped in clarifying the scope of the ontology (Liu et al., 2011)

1. **Data availability**: The Citizen Science ontology aims to integrate different datasets to ease sharing and solve non-interoperability issues in Citizen Science. With this regard, any Citizen Science projects worth considering should have most of it datasets available and accessible to the public. These criteria help in identifying the easiness in discovering Citizen Science projects based on the availability of the datasets.
2. **The popularity of the projects:** The number of participant in the projects: How widespread a project is, determines the number of citizens participating in that project. Therefore, in selecting the projects, the number of available projects were ranked based on the number of participants. The most participated project was then selected and considered with the other criteria. Moreover, there were few instances where few participants engage in a project, but the projects form the basis for an interesting subdomain worth considering. This value came to be after reading and releasing how a few platforms boast of the number of participants.
3. **The area of interest considered in the projects:** Projects serving the same purpose were not evaluated more than twice. The different and distinct goals of projects help in extending the scope to cover more areas of Citizen Science. An extended scope gives a clear depiction of the domain of Citizen Science.
4. **Structure and formats of the datasets:** The primary purpose of the Citizen Science ontology is built upon spatial relations. Therefore, most datasets and information to be considered in the knowledge acquisition state had a spatial component that comes with it. This spatial information can be in any format. Possible formats to be considered are CSV, XLXS, XLS, JSON, GEOJSON, XML and SHAPEFILES.

Table 5-1: Some Examples of the List of Projects Considered. Source: Author

	Project Name	**Domain**	**Link**
1	Geowiki	Land Resources (maps)	Link[20]
2	Did You Feel It? (DYFI)	Earthquakes	Link[21]
3	Atlas of Australian Birds	Birds	Link[22]
4	Anecdata	Botany, Entomology, water quality, Phenology, air and water quality etc.	Link[23]
5	Big Bug Hunt	Entomology	Link
6	Big Butterfly count	lepidopterology	Link
7	GeoTag-X	Disaster Risk Reduction	Link
8	iNaturalist	Biodiversity	Link
9	NatureWatch	Ice, Frogs etc.	Link
10	Mangrove Watch	Wetlands Monitoring	Link

[20] https://www.geo-wiki.org/
[21] https://earthquake.usgs.gov/data/dyfi/
[22] http://birdlife.org.au/projects/atlas-and-birdata
[23] https://www.anecdata.org/

Competency: Competency is an indication of the capacity of the tool in its usage. Therefore, the kind of knowledge to be captured in the ontology should be relevant in solving the problems at hand. The different datasets helped in identifying the types of algorithms and constraint-based search to be used. These competency issues are further discussed in the use case section of this projects.

Formality and Granularity: The domain analysis stage reveals that different subdomains can be formalised under Citizen Science. These Subdomains are distinct yet compatible when modelled together to complete the Citizen Science ontology. Considering the general characteristics of the reviewed projects, the following concept can be released to serve as the structure for defining the Upper-Level class knowledge in the Citizen Science ontology. Table 5-2 shows a general review of the groupings from the Citizen Science projects. The review in Table 5-2 explains the domain of interest and the intended purpose of the final Citizen Science ontology (Lozano-Tello & Gómez-Pérez, 2004).. The diversity of the domain of Citizen Science makes it more exciting and time-consuming to model the Citizen Science ontology. However, using an automated knowledge capturing tool was not an efficient means of knowledge capturing concepts and knowledge due to the underdeveloped domain of natural language processing (Ovchinnikova, 2012). Therefore, a manual means of grouping the domain of Citizen Science into deferent superclasses was adopted. The classes were grouped based on literature on requirements for performing Citizen Science projects discussed in chapter two.

Tools and algorithms: Tools and frameworks used are discussed in chapter four.

Table 5-2 Determining the Granularity and Formulation of the Ontology. Source: Author

Category	Knowledge to be Captured
Climate	Climates refer to the statistical averaging of the weather conditions over a period (UNFCCC, 2007). The study of climate and climate change are essential to understanding the drastic effect of our changing environment (IPCC, 2014). Therefore, numerous Citizen Science projects capture information on climate and climate-related issues. The aspect of climate to be considered is climate change. However, consideration was made to accommodate future work on this ontology to develop a more specific component of climate and weather-related issues.
Botany	Botany referred to the study of plants. Almost all lives on earth depend on plants from the biodiversity perspective. (Schooley, 2017). Therefore, numerous Citizen Science projects collect information on plants for research activities. Examples of these projects are the NatureWatch from Canada and the Inaturalist Citizen Science. The study of plant reveals several vital information and characteristics which form the basis for most life forms. There are many essential products which are carried out by plants and other organisms. This ontology considered the most aspect of plant information including photosynthesis.
Data/Time	Every Event occurs at an epoch. Time and dates are suitable means of serialising information that occurred at a specific point. Therefore, the concept of time helps in understanding the trend and occurrence of a phenomenon. It also serves as a means of classifying different climates and weather conditions. This ontology considers information of time and date as a means of serialising the different datasets obtained from the selected projects.
Classification (Biology)	Classification of species into groups due the common characteristic possessed by such species helps in identifying individual organisms on the earth. The science of grouping and naming organisms as a result of this unique characteristic is referred to biological

	classification. In the Citizen Science ontology, a detailed classification system was adapted to evaluate different types of species and to set relationships among these species.
Concepts	Concepts in this ontology refer to non-existing or intangible features that are ideologically and internationally accepted as a norm. Such concepts include standards and rules. This class definition helps in managing information about different datasets that have no geographic locations but has a unique standard about a unique geographic location.
Ecology	The study of an organism, interaction among organisms and their environments. The study of ecology helps in understanding most forms of relations that exist among organism in nature. This relation reveals unexposed potentials of the environments. It also provides means of capturing these potentials and utilising them for man's benefit.
Spatial information	Spatial information is the core basis for the design of the Citizen Science ontology. Information such as geometry, dimensions and topology was covered at this superclass. Most of the spatial knowledge used is the simple feature topological relations (OGC, 1999).

Step 2: Determining Data Sources

In Citizen Science, the information to be captured comes from heterogeneous sources, such as projects dataset, projects websites, domain experts, projects videos, manuals, field data among others. These set of resources were obtained from the evaluated Citizen Science projects. Table 5-3 gives a general overview of a list of datasets considered with their reference or sources.

There is an awe-inspiring number of data sources; it is vital to identify potential and appropriate data sources that can be used to capture meaningful Citizen Science domain knowledge. With these regards, the following data source reviewed in Table 5-3 gives an overview of the datasets. The following conditions were observed before formulating the Table 5-3.

1. **Adequate Information**: The dataset to be considered should have enough information to be captured in the ontology. There shouldn't be a case where the datasets only contain the location of observations and nothing else. Datasets with such limitations were not considered since the information to be captured from the note generated from the general public as part of the dataset.
2. **Geospatial Components**: Datasets considered for the design contains spatial information either in the form of metadata or specific geographic components in the datasets.
3. **Knowledge Coverage**: Domain expert were considered in an informal and unstructured interview. The domain of risk was considered, and knowledge regarding floods, earthquakes and landslides were considered as data sources from potential information to be used in the Knowledge acquisition stage.
4. **Clarity**: Clarity in the source of data determines how well-structured the data source could be. Therefore, data sources such as web pages selected should contain a clear and concise description of the information aimed at.

It is along these lines that helped in obtaining the required knowledge and significant learning and understanding through the information procurement process (Knowledge acquisition stage).

Table 5-3 Overview of the different Datasets Considered and used in the ontology design. Source: Author

	Datasets/data	Overview	Source
1	Did you Feel it (Datasets, webpage)	The dataset is a comprehensive catalogue of earthquakes resources from the general public. It comes in different datasets formats. These datasets contain instances of several observations as well as well-	(link)

		structured information in the form of metadata. The format selected for this dataset is Excel (CSV).	
2	GeoWiki (Datasets, webpage)	The Geowiki platform provides structure information on land validation resource obtained from the public. The datasets are mostly Images and Shapefiles. A preview of the data shows a list of terms that describe land resources. Much of it is considered in the Knowledge acquisition stage.	(Link)
3	NatureWatch (Datasets)	NatureWatch programs dataset can be selected according to province and date range. Data are presented in CSV format. It contains adequate information to be captured at the knowledge acquisition stage. The datasets are in four different categories. Frogs, Plants, Ice, Worm and Milkweeds.	(link)
4	Big Butterfly Count (Webpage)	The webpage describes the results of the butterfly count in quite an exciting way. It indicates most of the concepts and action is taken to realises the said dataset. It will be a potential source of information at the knowledge acquisition stage.	(link)
5	Anecdata (Datasets and webpage)	Anecdata is an online science repository for any person who wants to assemble or offer normal data on the environment. The dataset cut across different fields of environmental datasets generated by the public. The datasets are in different formats ranging from CSV to geoJSON including XML.	(link)

Step 3 Knowledge Acquisition

The Knowledge acquisition is the act of capturing information from the data source into the ontology. At this section, the knowledge acquisition approached is based on the Knowledge elucidation proposed in the generic ontology development framework. The Knowledge extraction considered three (3) different phase. The result from each phase is expressed in a formal ontology language (owl). The three-phase includes a survey of existing ontologies that express subdomain knowledge in the domain of citizen, Unstructured discussions with a domain expert to capture knowledge and the conversion of the required data to owl using appropriate tools. At the conversion of project data to owl three different tools were considered. These tools are the Cellfie plugin come with protégé, JSON_2_OWL tool re-edited from GitHub platform and a TEXT_2_OWL tool developed by the author for converting Web page and pdf document to owl files in the ontology. Each of these tools is explained in the appropriate section.

Reuse of Existing Ontologies: *(Existing Ontologies)* The following ontologies are evaluated to be used in the Citizen Science ontology due to the various domain of Citizen Science. With much emphasis on the reuse component of the selected methodology, each ontology was selected as a result of the knowledge and field/ domain they capture. Table 5-4 shows some selected existing ontologies that were reviewed and used in the Citizen Science ontology.

Table 5-4: Different classes and concepts selected and used in the design process (Emphasising the Reuse Component of the Selected Criteria in Section 3.4). Source: Author

Ontology	Number of classes	Area of interest

Plant ontology	30	Plant descriptions with plant types and plant names.
Vertebrate Taxonomy Ontology	25	The descriptions of vertebrates with types and names.
Social Insect behaviour ontology	25	Insect behaviour such and insect locations.
BBC Wildlife ontology	All	Animal behaviour and animal classification.
Ordinance Survey spatial relation	All	Spatial relations to map spatial entities.
W3C Geo vocabulary	All	Encoding of spatial information based on the WGS84 coordinate system.

Converting Data to OWL Ontologies

Converting Excel to OWL (EXCEL_2_OWL): The conversion of excel data to OWL was done with the cellfie plugin tool implemented in Protégé. Several excel formats are converted to the latest version of Excel (Xlxs). Examples of the formats converted are CSV, Xlx and dBASE Table in Shapefiles. The cellfie plugin tool permits mapping rules to be developed base on the Manchester OWL syntax. The syntax helps in structuring and selecting different concept in the excel to represent classes or object properties. The rule formulation tab was selected, and the rules established were assigned to the imported data. Different datatypes exist in Excel; these datatypes consist of all primitive datatypes as well as non-primitive datatypes. Classes, subclasses, annotation and data properties were selected based on the semantics of the data and the datatypes in the data. Figure 0-4 in the Appendix shows an example of the datasets imported into protégé with the cellfie plugin and the rule definition base on the Manchester OWL syntax. From Figure 0-4, there were 16 columns in the data, columns like Scientific name and local name were assigned the same as relations. The overall process at this stage in the importation is shown Figure 5-5. As displayed in Figure 5-5, all the different formats were manually converted to the XLXS format of Excel. Mapping rules were obtained based on the natural clustering in the datasets. The datasets were imported to the with the plugin into the protégé environment and were stored internally or semantic structuring and alignment. This link points to the source of the plugin.

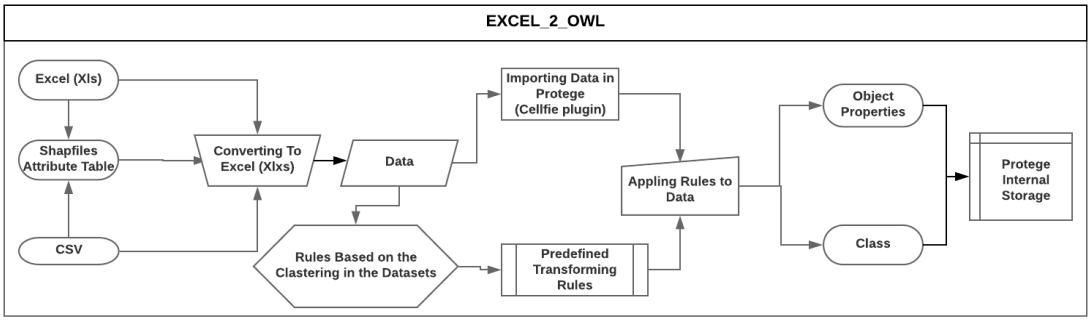

Figure 5-5: Overall Process for Converting Text to OWL. Source: Author

Converting JSON to OWL (JSON_2_OWL): *The* JSON to OWL tool is a simple JSON converter, developed by structuring JSON and GEOJSON files with the OWL syntax. JSON is a lightweight data-interchange format structured like text files which allow machines to parse and generate easily. The tool was developed with JavaScript language and Ajax framework. The comparison between JSON syntax and OWL syntax alluded a simple mapping strategy that allows the JSON semantic containing concepts, properties,

constraints and values directly mapped to the OWL ontology. Figure 5-6 shows the flowchart of the JSON_2_OWL converter. Different data formats such as GEOJSON, XML and WKT that can easily be converted to JSON were also manually flattened to JSON and converted to the OWL ontology. The tool has an HTML page that allows easy upload of JSON files. The uploaded JSON was parsed to enable the key-value pair in JSON to into triples in the OWL ontology using both spatial and non-spatial relations.. The tool was obtained from this link and modified by the author.

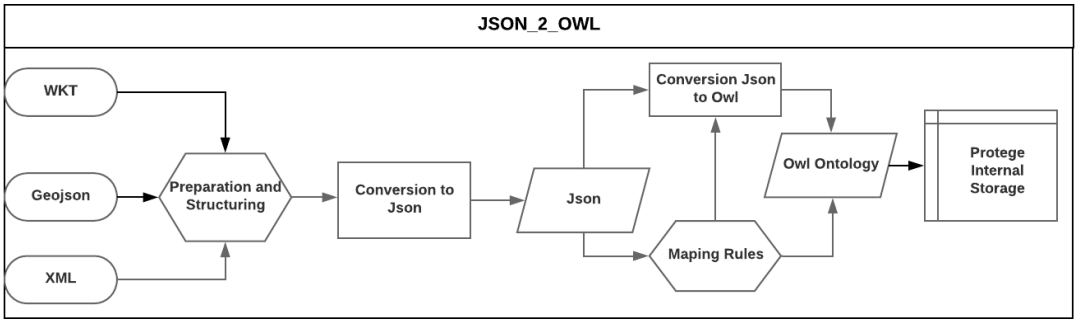

Figure 5-6: Overall Process for Converting JSON to OWL. Source: Author

Converting Text to OWL (TEXT_2_OWL): *The* Text_2_OWL tool was developed to select keywords that exist on saved web pages and pdf files to OWL ontologies. The tool is developed to work on text files only. Therefore web pages and pdf documents are manually converted to text files with the 'txt' file extension. The translated text files are always unstructured with more unwanted characters. The un-structure file is first structured to remove unwanted characters like space and words such as on, is, with, and, an etc. The words in the structured text file are then ranked based on the frequency of occurrence, and the most occurring words are converted to the OWL ontology. The final OWL ontology is edited to remove unwanted and duplicated words. Figure 5-7 shows the workflow for the conversion from text to OWL.

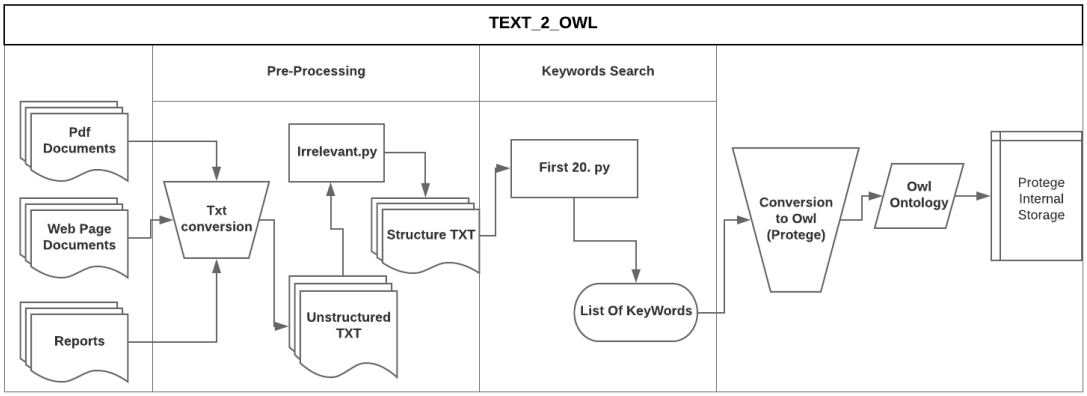

Figure 5-7: Overall Process for Converting Text to OWL. Source: Author

All the generated information was stored in the Protégé internal memory for the development of the ontology at the next section.

5.3.2. Ontology Development Phases (B)

This section describes the definition of classes and mapping of class-subclass hierarchy and their relationships. The development phase consists of coding and design of the ontology. A list of relations which includes spatial prepositions was selected as object properties for the design of the ontology as discussed in chapter four. This section in this report considers most of the Upper-classes **Knowledge** and

Data in the design. However, there are cases where concepts of other Upper-Levels Classes maybe mentioned. Figure 5-8 shows a general overview of the steps considered at this stage.

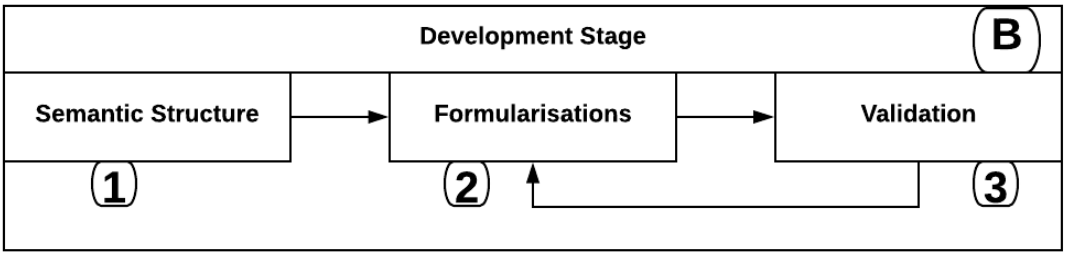

Figure 5-8:General Overview of the Ontology Development Stage. Source: Author

The upper-level class **Knowledge** is in relation to several classes which include climate, Nature, Science, Classification, Concepts and Botany. Each subclass has other subclasses which include Classification, Risks, Date, Food, Habitat, interaction and other. These categorisations are to depict the diverse range of Citizen Science domain. The processes of organising the knowledge acquired from the knowledge acquisition stage stored in the internal memory are used in the formularisation stage based on the semantic structure of the ontology. The expected validating and representation of the knowledge, creating and operating inference mechanisms, and dealing with uncertainty are based on the semantic structure of the ontology. The section is divided into three sections: Semantic Structure which gives a general model for the design, the Formularization which gives examples of how knowledge was captured and aligned based on the selected semantic structure and validation which discusses means of confirming the correctness of the captured knowledge. Moreover, almost all the processes are iterative throughout the design.

Step 1: Semantic Structure

The act of identifying and analysing the knowledge available in the acquisition stage forms the critical elements of knowledge capturing in the ontology design. The semantic structure serves as a way of ensuring a logical flow of defining concepts in the ontology design. The semantic model for the Citizen Science ontology extends from a simple class-subclass hierarchy to a more multifaceted model. The theory used to map, arrange and organised the stored internal OWL ontology in the protégé environment is the model proposed by the Generic Ontology Development Framework to classify ontological classes and subclass hierarchy. The model is as follows

$$O_i = \{C_i, C_{isSubClass}, Rel_{Ci \rightarrow cj}\, Instance_{ci}\, Rules_{oi}, Axioms_{oi}\}$$

This model gives an impression of what constitutes a class, subclass, relationship, instance, rules and axioms in the ontology. The meaning of the terms used and how there are checked in correspondence to ontological knowledge are as follows.

Ontology, (O_i): This is the ontology for a specific Subdomain considered as superclasses in the Citizen Science ontology. The seven upper-classes serves as seven subdomain ontologies combined to form the Citizen Science ontology. The scope of this section considered only the upper-level knowledge as a sub-ontology.

Class, (Ci): Classes represent a group of concepts the is adequate to describe and help formulise some specific knowledge in the subdomain. Classes give a deliberation system for gathering assets with corresponding attributes. There are several pieces of concepts that are captured in the ontology as classes. In the ontology, 15 supper classes were obtained as a result of the natural groupings that occurred in the dataset (acquired knowledge). This 15 superclasses cut across almost all field in the Citizen Science domain. Details of this classes are discussed in the document specification section. Examples include Hazards which is a subclass of the class Nature, Plants which is a subclass of the superclass Botany. The concept Plant was obtained from the Plant Ontology (PO) analysed for reuse. Due to the granularity and scope of the Citizen Science ontology, other existing ontologies are considered as classes or subclass of the semantic structuring.

Some fundamental concepts in the classification of biological species were adopted. A classification such as the Class "Life" with Subclasses "KingdomAnimalia", "KingdomPlantae" and the other kingdoms to represent the phylogenetic tree of life. A simplified representation of living organisms that are classified according to their characteristic in nature.

Subclass, ($C_{isSubClass}$): An adequate description of the specific portrayal of a class is expressed in their subclasses. There are different levels of subclasses in the ontology due to the complexity of some domains as compared to other domains. For example, the class Hazards has subclass Earthquakes, floods, Landslides. Subclass earthquake has subclasses representing more specific concepts such as EathquakeCause, EathquakeTypes and EathquakeOccurrences. Examples in the fauna world, the class Chordata has subclass Aves class Animalia. The general properties of a class are inherent in it subclasses. That is if a class B is a Subclass-Of a class A, all instance of class B are instances of class A. The acquired knowledge in the Knowledge acquisition stage has more than 300 subclass-subclass hierarchy. Examples include the class Risk has subclasses Earthquakes which is also a subclass of the class Natural hazards. Subclass EaethquakeStation of the class Earthquake has subclasses EarthquakeIntensity. Moreover, the classes and subclasses in this ontology are categorised by specifying their attributes and associations among them.

Relationships ($Rel_{Ci \rightarrow cj}$): Relations among classes and individuals exist in this ontology, and are means of associating classes as well as individuals. In the OWL syntax, relations are termed properties which maps two classes or individuals and data values. Two types of properties were recognised in this ontology. These are the object properties and data properties. Object properties link two or more individuals in a class or among classes and data properties assign data values to individuals/instances. Examples of relationships expressed in the Citizen Science ontologies are data properties: HasID and HasName, Object properties: RecordedAt. Expression: earthquake OccursAt Station and Station HasID: integer and HasName: String.

Instance or individuals (Instance): Individuals or instances are the least level objects of a class in ontologies. Classes have a set of individuals that exhibit the properties of that class, and all instances fall under the OWL:Thing. Many facts were indicating membership of certain classes to other classes.

Rules ($Rules_{oi}$): Rules are a well-known practice that describes logical inferences obtained from the assertion of a specific nature. It offers high-level of expressiveness such as constructors for composite properties. Rules are mostly developed in a logical programming language which supports logical reasoning. The protégé SWRL tab is used to defined rules under the semantic structure of the Citizen Science ontology.

Axioms ($Axioms_{oi}$): Class axioms help express class descriptions which form the fundamental element for defining classes. In general, axioms describe a well-known fact that exists in a domain with logical assertion statements.

Selecting classes and the type of relations (predicates) that maps a class to another class depends on the information contained in the datasets converted to the OWL. The ontology formularisation defines some class-subclass hierarchy in the ontology.

Step 2: Ontology Formularisation

Semantic association of the different sections and different knowledge contained in the generated OWL ontologies across all resources were not the same. Therefore, the class-subclass hierarchy and the relationships between different classes and entities were defined manually. The manual strategy is to provide more logical, coherent and consistent semantic association among concepts. This section described some class-subclass hierarchy defined and designed to answer the competency questions in chapter four. Most of the concepts discussed here are a subclass of the superclass nature defined as a concept for both domain

specific and generic concepts in the ontology. The class prefixes are omitted for clarity and simplicity. Words in the form **Grass** represent classes, and ***HasType*** represents relations or data properties.

Considering the Use Case for Integrating Biodiversity Conservation into City Conservational Planning

1. In the biodiversity domain, more classes were created. Considering the integration of biodiversity conservation into city conservation planning, two top-level classes **Nature** and **Concepts** were created. The purpose of the superclass **Nature** is to capture the knowledge in the domain of both biodiversity and landscape models which includes LandUse and **LandCover** information. Whiles **Concepts** is to define a conceptual framework for modelling data during the data integration process.
2. The class land is a subclass of Nature which has a subclasses **LandUse** and **LandCover**.
3. **LandCover HasTypes Grass**, **Asphalt**, **Trees**, **BareGround**, **Water**.
4. **Grass** falls under class **Plant** and has ***SubClasses*** **Wheat**, **Oat**, **Teff** and others.
5. **LandUse** also ***HasTypes*** **Recreational** (**Parks** ***Is-A-Type-Of*** **Recreation**), **Transport** (**Roads**), **Agriculture** (**Farmlands**), **Residential** (**Housing**), **Commercial** etc. **Agriculture** includes **Grass**.
6. The axiom **Grass** ***Is-A-Type-Of*** **Food** is realised from this classification. This classification and grouping style enabled the capturing of knowledge in the form of graphs connecting land use and land cover information to the concept Nature.
7. Concept **Nature** has subclasses **LivingThings** and **NonLinvngThings**. These two classes are to enable the ability to extend the ontology to capture all knowledge needed depending on the purpose of the data to be integrated. This Superclass has different subclasses such as **Plant** and **Animals**.
8. Class **Plant** captures knowledge in flora world and **Animal** captures knowledge in the fauna world. Under listing the class **Plant** reviews different characteristics of **Plants** ranging from **PlantTypes**, **PlantDevelopmentalStages** to **PlantsUsage**. Most of these characteristics are obtained from the Plant Ontology (PO).
9. The class **Animal** expresses knowledge of different animals. It however considered potential classes from both the BBC Wildlife Ontology and the Geospecies Ontology.
10. A subclass is the class **Insects**. Class **Insect** expresses the types, social behaviour and different examples of insects. Most of these knowledge were obtained from resued classes from the Social Insect Behavior ontology (SIBO).
11. Biodiversity domain reviews several axioms such as **Garter Snakes** ***AreLocatedIn*** **North-America**. This and other axioms present several indications of plants location using the ***FoundIn*** property defined in the ontology and assigning specific species to their frequently reported locations (Counties). Therefore, the class **Reptile** has ***SubClass*** **Garter Snakes**, and they are located in class **North-America** which is a ***SubClassOf*** **Country**. Moreover, the ontology communicates less information on the different individuals under the different classes. It, however, presents the notion of the individuals as expressed in their respective classes.

Formularisation Considering Spatial Information

In dealing with geographic locations, the Basic Geo (WGS84 lat/long) vocabulary developed by W3C Semantic Web Interest Group (SWIG) and the Ordinance Survey Spatial Relation Ontology were considered in the reuse section.

1. The Basic Geo Vocabulary introduces a concept called **SpatialThing** which expresses the relation and entities adequate for modelling and encoding spatial information in the ontology. It has several attributes such as ***geo:lat*** and **geo:long** describing both the geographic latitude and longitude of a point in space.
2. This relation was directly imported and assigned with the ***SameAs*** relation to the ***HasLat*** and ***HasLong*** relations respectively in the ontology. The basic GEO vocabulary ontology is used for handling geospatial data according to the WGS-84 geodetic reference system. Under the Concept class formulation, a more proactive prosses of describing the location of an object, the physical

composition of the object in question and other significance of the objects under study were considered. Therefore, spatial objects like Parks which is a subclass of Recreation were considered to be within a city's boundaries.

3. A **City** has both **LandUse** and **LandCover** that determines the required conservation plan.
4. Spatial relations serving as predicate among these spatial objects (Thing) were used. Examples include Cities ***Contains*** **LandUse** such as **Recreation** and **Transport**. (**City** ***Contains*** **Roads**).
5. The class spatial object is a ***SubClassOf*** **OWLThing** but not considered as Upper-Level Concepts because most of the Upper-Level Concept uses Subclasses of Spatial Objects and relations to map two or more objects. **SpatialObject** ***EquivalentTo*** **SpatialThing** has two subclasses **SpatialFeatures** and **SpatialGeometry**. The definition of spatial features and geometries and their properties follows the Ordinance survey spatial relation ontology.
6. The ontology formularisation in the development stage is based on the selected semantic structure. At this stage, the merging of the selected class with their object properties was aligned based on the semantics and common vocabulary of the domain of Citizen Science. From the ontology reuse section, Selected classes from the previewed and adopted ontologies were merged and added to the ontology. An example is the following, **EarthquakeLocation** and **EarthquakeEpoch** considered under the class Earthquake were merged into a more generic representation **EarthquakeStations,** which falls under '**Earthquake'** a ***SubClassOf*** **NaturalHazards**.

Formularisation Considering Use Case Set 2

When one wants to assess wetlands conditions for wetland birds and validation of GeoWiki Land Use and landcover information, the datasets necessary to be integrated will be extensive combined information on most different species. Such species include waterfowls, insect, rodent datasets. The generated OWL from the knowledge acquisition stage was organised as follows. The waterfowl and insect were formulated using the relation diets from the A-Z animal information.

1. The triple '**Waterfowls** ***eat*** **Insects'** such as **Dragonfly** was created.
2. **DragonflyNymphs** ***GrowOn*** **Wetlands**. **Waterfowl** ***FoundOn*** **Wetlands**.
3. Insect lay their eggs on **WaterlogsAreas** which give their eggs nurturing condition for proper growth.
4. The class **Insect** and **Birds** in the ontology were mapped with the relation ***Eat***.
5. However, under the Class **Insect,** there is the subclass **DevelopmentStage** which ***HasType*** **Nymph** as an instance '**Nymph** ***GrowsOn*** **Wetlands'** serves as a data property in the ontology for the class **Nymph**.
6. The notion **Swamps** ***Breed*** **Insects** describes possible locations where **Insects** can survive in **Wetlands**. This mappings and relations can be used to combined datasets on insect and waterfowl.
7. **Wetland** ***SubClassOf*** **LandCover** ***Inhabits*** **Waterfowl** ***SubclassOf*** **Birds** (N-ary triple formulation). Therefore, multiple sightings of the waterfowl birds in a particular location can be said to be a Wetland.
8. **ShallowWateryAreas** ***Is-A-SubClass-Of*** **Wetlands** which includes **Swamps.** This is mapped with the relation **Swamp** ***Is-A-Type-Of*** **ShallowWateryAreas**.
9. Another triple is **OWL** ***FoundIn*** **Forest**: This relation provides the notion that most of the Owl sightings are found in Forest areas. Therefore, Forest land cover types submitted by users for that location can be cross-checked if there are a cluster of Owl sightings at that particular location.
10. **Swamp** ***Is-A*** **ShallowWateryAreas** and **Wetlands** ***Is-A*** **ShallowWateryAreas**: The two subclasses of shallow watery areas can aid in validating Wetland land class on the GeoWiki platform.
11. **Dragonfly** ***HasDevelopmentStage*** **Nymph** ***FoundOn*** **Wetlands**. In the life cycle of the insects, there exists a nymph stage. This stage is mostly on wetlands. From the dataset, dragonfly Nymph is found in wetlands. Therefore, these relations in the ontology and other information can be used to validate the land classes of such locations form GeoWiki user input.
12. **Grass** ***SubClassOf*** **Plant** ***FoundOn*** **Grasslands** are possible integration strategies adapted during this formularisation stage.

Formularisation Considering Use Case SET 3

The Use Case "Developing Vertical Forest", vertical forest can exist at different height levels. Information on different animals for specific height was handy.

1. Under the class, **Domain** has ***SubClass*** **Ecology. Ecology** has ***SubClass*** **Food** which describes different diet for different species. From the A-Z Animal Information, the relationship diet (***Eat***) and Inverse relation ***EatenBy*** was used to map the two concepts. Examples include the triple "**Birds** ***Eat*** **Insect**" and "**Nectar** ***IsConsumedBy*** **Nectarivore**" (Nectar sucking birds). Moreover, nectarivore has mixed diet which includes both insect and nectar. The relation groups different bird species and plants at different height levels due to their characteristics at different height levels.

Step 3: Ontology Validation

An ontology validation checks the accuracy of the formularisation for the intended purpose and domain. The ontology validation at this stage checks if the Citizen Science ontology captures most of the knowledge required information needed to integrate the selected datasets at the pre-development stage. As proposed by the generic ontology development framework, this stage considered the data-driven approach for validating the ontology.

After the full mapping and merging of the generated OWL ontology, the reasoner was initialised to check for consistency. Contradictory concepts were resolved manually in accordance with the OWL 2 syntax. The final consistent and uncontradicted ontology was converted into OWL DL for reasoning. The formularisation under the protégé environment allows conversion of the generated ontology to different forms of formal ontology language such as RDF, OWL FULL and any other type of machine-readable formats. The formularisation process also considered the enrichment of class definition base on class attributes and class-attribute-slot-value-type. Examples include strings, Booleans and others as well as slot cardinality. More classes seem similar on the formularisation stage. Such two-similar concept or classes were aligned and merged or disjointed depending on the semantics of the two classes.

5.3.3. Post Development Stage

Ontology Documentation

The ontology documentation is a crucial step in the life cycle of ontology development. It is necessary to provide correct documentation of a new ontology to facilitate precise and correct interpretation of the semantics expressed in the ontology (Meaning of the classes). This enhances the use and clarity of the ontology structure for diverse groups of users. More importantly, all the assumptions that were made while developing an ontology are written explicitly in a natural language to avoid misinterpretation of the logic. The documentation of classes and their intended means is shown in Appendix 1 (The glossary of terms).

The Designed Ontology

The final designed ontology comprises a list of concepts expressed in the formal ontology language to describe concepts in Citizen Science. The ontology is internally stored in the developing environment of protégé ontology editor. Figure 5-9 shows a preview of the developed ontology in the protégé environment. Individuals were not emphasised during the design stage. Generating individuals were set to take place during the data integration stage in Chapter 6. This is to make sure the ontology can be used to map different sightings of species and other relevant information ad instances of their respective class. However, the mappings among few individuals were conducted based on the relationships that exist among classes. Figure 5-10 and Figure 5-11 show the general overview of how ontological classes and relations are mapped both among classes and individuals in the ontology. Individuals are made to inherently subsume the properties of the class they belong. Individuals are mapped with the object properties connecting the different classes.

Therefore, individuals are better understood as equivalent members of classes in the ontology. The designed ontology has few individuals describing some specific aspect as well as clarifying the notion of individuals. Most individuals are obtained during the modelling of data for the integration purposes. Classes are considered to portray characteristics of individuals. Data column and cells in the datasets communicating the semantics of a class in the ontology are assigned with the semantic type of that class as individuals in the column. These individuals can be mapped to other individuals in different classes using the relation existing among the two classes. An example is in Figure 5-11, where OWL 1 is mapped to Drag 2 using the relation *Eat*. However, before the mapping can be validated, a check is made to the data properties of the two individuals. If some common and relevant data properties among the two individuals share the same value, then the two individuals in the classes can be affirmed to have the proposed ontological relation *Eat*. From Figure 11, there are three different classes in the ontology. These three classes are mapped with relations among each other. The relations are considered to be inherently part of all individuals of the class. However, depending on the datatypes in each sighting and the information in the data columns from the datasets, the individuals can be mapped to each other by considering the datatypes for each data property relevant for the specific class. An example is the triple **Dragonfly** ***HasHabitat*** **Forest**. Individuals of the class **Dragonfly** are assigned different subclasses of the class forest depending on the information contained in the dataset. Drag 2 is assigned the subclass **Deciduous** due to the sighting notes in the datasets.

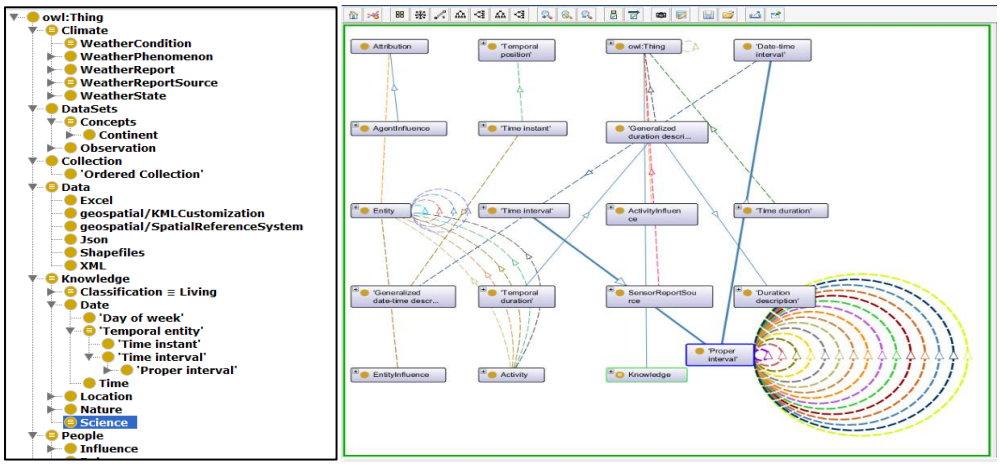

Figure 5-9: Preview of a Section of the Designed Ontology. Source: Author

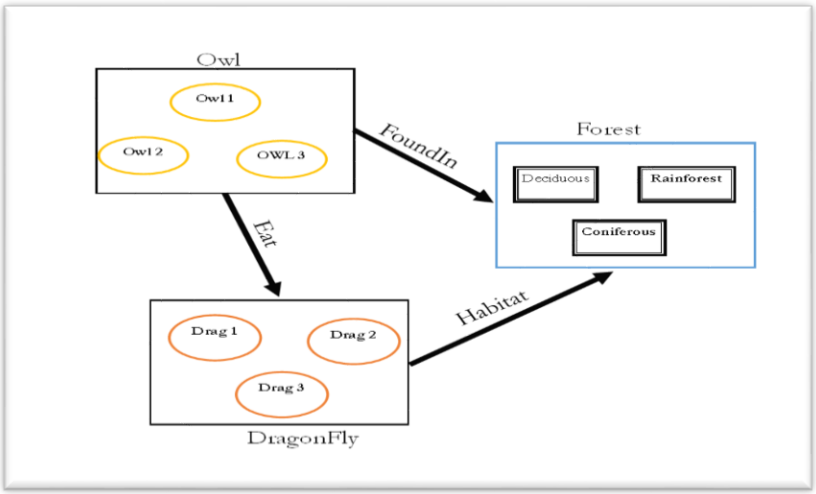

Figure 5-10: Preview of a Section of the Designed Ontology. Source: Author

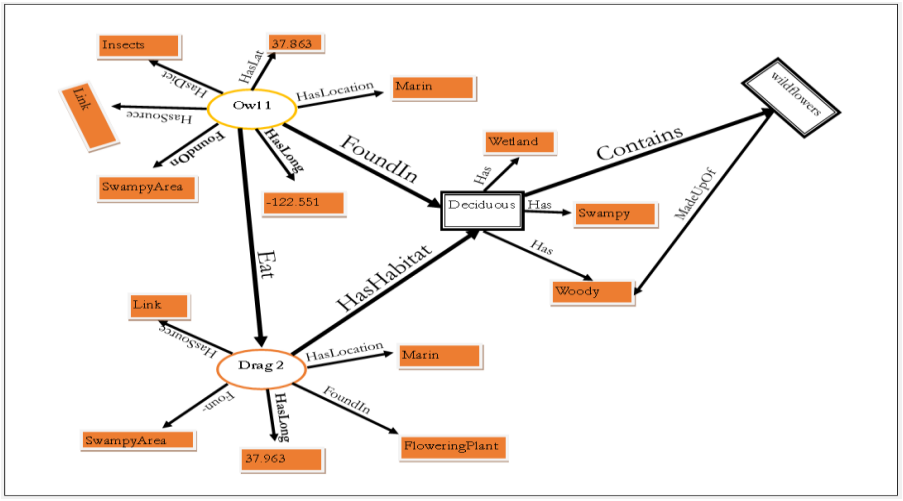

Figure 5-11: Mapping Individuals using Both Object and Data Properties. Source: Author

5.4. Supporting Activities

This section gives an overview of the implementation strategy for the designed Citizen Science ontology.

Implementation

The task of implementing the ontology in an ontology language required an environment that supports the ontology editing and reasoning. Features considered for the selection of the environment include the following:

1. A lexical and syntactic analyser to guarantee the absence of lexical and syntactic errors.
2. An editor for adding, modifying, and removing definitions
3. A browser for inspecting the library of ontologies and their definitions
4. A searcher for looking for the most appropriate definitions.
5. Evaluators for detecting incompleteness, inconsistencies, and redundant knowledge
6. An automatic maintainer for managing the inclusion, removal, or modification of existing definitions

Together with the implementation, the information about the ontology gathered the made the ontology implementation straightforward. The process described in Section 4.3.2 was formalised into a formal model of an ontology language. The Citizen Science ontology formularisation was done using the Protégé ontology editing environment together with the HelmiT 2 Reasoner.
After exhaustive analysis and structuring in the previous sections, the step of implementing the ontology has become a straight-forward task. The Citizen Science ontology is implemented in OWL using Protégé 4.3 together with the HelmiT 2 Reasoner. The HermiT 2 is an ontology performance tool that uses a set of patterns to find possible performance problems in an OWL ontology. HermiT has been used intensively to ensure it does not report any problems that could affect reasoning performance.
The following emphases were made on the implementation stage of the designed Citizen Science ontology.

1. All element with geographic information were conceptualised using axioms, classes and relations under **SpatialThing** Class.
2. Spatial Relations were considered as Objects properties describing both **DomainSpecific** Concepts and **GeneralConcepts** (Subclasses Under the Domain class)
3. All spatial Relations used in the ontology were duplicated due to their use. The duplicates represent attributes relations and spatially defined concepts (Based on the nine-intersection model in Table 4-2.

6. DATA INTEGRATION, QUALITY TESTING AND DEPLOYMENT

The act of using an ontology as a surrogate for the semantics of a domain has never been natural in ontological engineering. However, the Citizen Science ontology has been designed to enable two or more systems (dataset) to be compatible with each other and to increase the sharing of such information on the semantic web. This section describes the integration process, quality testing, ontology maintenance and deployment strategies for the designed ontology.

6.1. Data Integration

The testing of the ontology as a proof of concepts is to use the ontology for modelling different dataset to make them compatible. This section describes the integration of citizen-generated dataset based on the inherent properties contained in the dataset, the schema and semantics of the designed ontology. The integration process is composed of a review of different integration tools and modelling of the different datasets using the ontology as a surrogate within the selected Integration tool.

6.1.1. Selecting an integration tool

The process of modelling the different datasets based on the semantics and the schema of the ontology requires a platform for the integration. This section presents some potential integration tools that were considered. These integration tools include the Karma Data Integration Tool, The Talend Integration Tool and the Karma Ontology Mapper. The following criteria were considered during the selection process.

a. Availability of technical support
b. Potential Capabilities and Technical support
c. Efficiency and speed during usage
d. Adequate documentation
e. Buds and Debugging possibilities
f. Searching Capabilities
g. Analyzer
h. Modifications ability

A detailed comparison and analysis of the different tools can be found in Table 0-5 in the Appendix. Applying these criteria to the three common data integration tools above, the Karma Data Integration Tool was more appropriate. The selection of the different tools and the different quality indicators served as the user requirements for modelling different datasets with the ontology (User Requirement for using the ontology on the integration platform).

6.1.2. Modelling Data with the Ontology in the Karma Data Integration Tool

The karma tool is an open source program that allows designers to integrate different datasets base on an ontology schema and semantics. The University of South California developed it for integrating datasets of different form and format. To efficiently use the karma tools, knowledge of **Maven 3.0** and **Java 1.7** platforms are required. Full installation and configuration can be obtained at the Karma Website[24]. The overall process for modelling and integrating the datasets are shown in Figure 6-1.

[24] https://github.com/usc-isi-i2/Web-Karma/wiki

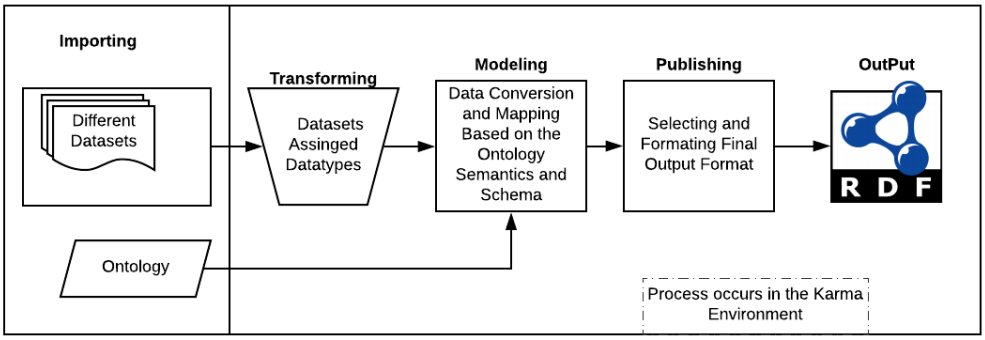

Figure 6-1: Overall Process for Converting Different Datasets. Source Author

Step 1: Importing Data

The designed Citizen Science ontology was first loaded into the Karma environment to serve as the base model for transforming the different datasets. Using the import tab, Karma can import data from both structured and unstructured data. The different datasets for the use cases were loaded into the ontology. These datasets had no field or column in common but had some inherent properties that can be integrated using the semantics and the schema of the ontology. Figure 0-7 shows an example of the data importing from Excel. Details on how to import different datasets from different platforms can be obtained at the karma data import[25] website. However, most of the datasets contain spatial information. Therefore, the spatial data[26] import tab was also used.

Step 2: Modelling and Transforming the Imported Datasets

After importing the different datasets into the karma environment, the Karma tool provided different commands that used the semantics of the ontology to transform the datasets into triples. The different datasets were transformed by assigning different cells and columns to different classes in the ontology. Table 0-3 shows a detailed connecting strategy for the different datasets using the properties defined in the ontology. This transformation is based on the inherent semantic properties in the ontology as compare to the datasets to be modelled. Details on the various commands to transform the datasets can be found in the data transformation manual[27] on the karma website. An example is a dataset containing plants sightings; it had 30 columns and 19800 rows. This dataset was moulded by first assigning the latitude and longitude columns to the HasLat and the HasLong properties in the ontology. The two properties were mapped to a point class which has a SameAs relation with the different plant species. Mapping for both Latitude and longitude were considered for **Point** under **Geometry**. The following triples in the ontology were considered during the modelling and transformation stage. Example of the mappings used includes **Owl *FowndIn* County**, where the different sightings of Owl were assigned to their respective county as reported by the datasets. Owl *FoundOn* Forest Area, Where the different species of Owls reported in the datasets were assigned to their respective habitats. **Owl *FoundOn* SwampyArea, Owl *Eat* Dragonfly *FoundOn* SwampyArea, Sunbird *Is-A-type-Of* Bird, Sunbird *Consumes* Nectar; Swamps *Breeds* DragonflyNymph** and **Waterfowl *Found*On LandUseType.** The column city was assigned spatial feature, and the relation C***ontains*** were used to set the relationship between the feature location and feature geometry. Columns like species name and species genus were assigned to their classes Species and Genus

[25] https://github.com/usc-isi-i2/Web-Karma/wiki/Importing-Data
[26] https://github.com/usc-isi-i2/Web-Karma/wiki/Working-with-geospatial-data
[27] https://github.com/usc-isi-i2/Web-Karma/wiki/Transforming-Data

respectively. Figure 6-2 shows a transformation and modelling of a dataset. In Figure 6-2, the columns city and Country was mapped with the using the triple **City** ***contains*** **County**. Some Land use classes were modelled by using a spatial relation contains. Such information includes Swamps *Breeds* Insects. Therefore, swamps contain conditions necessary for breeding nymphs of insects. An illustration is shown in Figure 6-2.

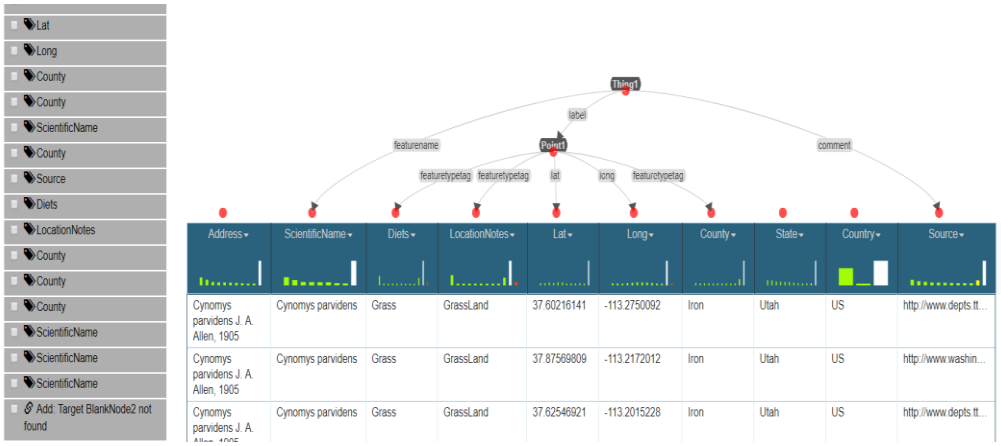

Figure 6-2: Modelling and Transforming the Imported Datasets: Source Author

Step 3: Publishing the modelled dataset in RDF

Publishing the data on the karma platform has several advantages since different data format can be published and used from the combined datasets. RDF triples were selected for easy integration into the designed ontology if the need arises. The selection of the RDF triples was to ensure adequate querying and to test the compatibility and interoperability using SPARQL query. Figure 6-3 shows a preview of parts of the triples (The: u_b50 represent different species contained in the different dataset). Moreover, the two images represent different visualisations styles using the c[28]. The RDF data format gave several options for querying the data for the necessary information. Some of these options include adding it to the designed ontology or querying it as a separate file through a python environment. Due to simplicity, efficiency and smooth querying, the generated triples were queried through a python environment using a translated competency questions. The next section presents the query of the RDF dataset.

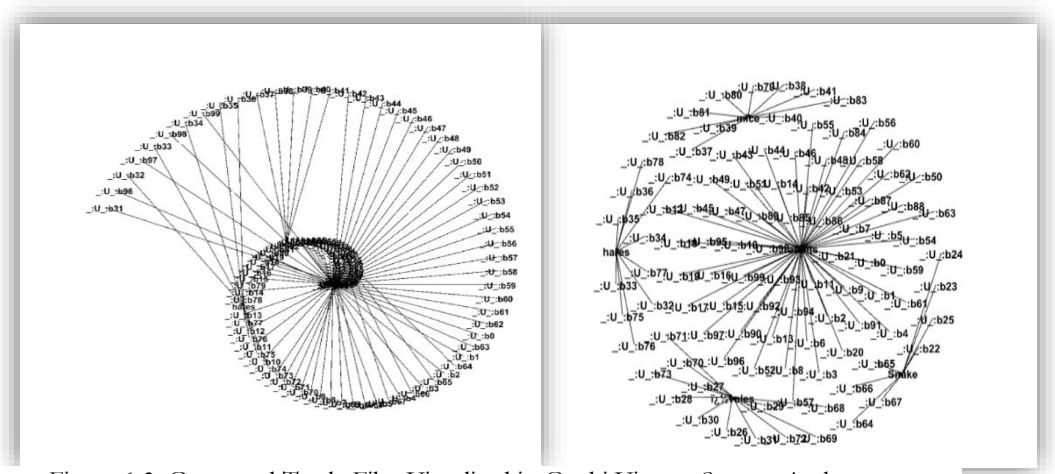

Figure 6-3: Generated Turtle Files Visualised in Gephi Viewer. Source: Author

[28] Gephi is the leading opensource visualization and exploration software for all kinds of graphs and networks.

6.2. Quality Testing

Evaluating the designed ontology can be considered as means of testing the quality and capabilities of the ontology. Ontology evaluation helps in determining its potential capabilities and importance. This section used two different means of assessing the quality of the generated Citizen Science ontology. The two processes include using SPARQL query to test the integrated dataset (section 6.1.2) and using a well-developed metrics from the semiotic theory (Semiotic Metric Suite).

6.2.1. Use Case

The use cases testing was done to confirm the integration capability of the designed ontology. It followed a two-step process.

Step 1: Translating Competency Questions to Queries

The procedure for creating SPARQL queries from the competency questions in Section 4.3 followed the paradigm of set and logic theory (Avraham et al. 2013; Schwartz et al. 2011). The resulting SPARQL queries for the use case and the reasons for connecting the different classes are shown in Table 0-3 in the appendix. Table 6-1 aims at giving an overview of the first three competency questions in the dataset. Reviewing Table 6-1 shows that the natural language semantics in the competency question can be transformed into SPARQL queries using set and logic notation. The prefix for the ontology is denoted with "cs"

Table 6-1: The First Three Competency Questions (Q1.1, Q1.2 and Q1.3) with their Corresponding Queries to Test the Quality of the Ontology. Source: Author

Question number (label is provided as [QSet. Question no])	Query
Q1.1 Which region is the highest reported number of species?	SELECT (?o as COUNTY) (COUNT(?o) as ?No_of Sightings) WHERE { ?s cs:FoundIn ?o . } GROUP BY ?o ORDER BY DESC (?No_Sightings) Limit 3
Q1.2 What are the land use and land cover characteristics of those regions?	SELECT (DISTINCT ?o as ?Land_Information) WHERE { ?s ?p "San Francisco" . ?s ? cs:FoundO . }
Q1.3 Are there any risks of natural disaster reported in these areas?	SELECT (DISTINCT ?o as LandHazard) WHERE { ?s ?p "San Francisco" . ?s cs:HasRisk ?o . }

Step 2: Testing queries on the combined data

The quality assessment is to test if the generated model contains information from the different datasets. The generated queries from the competency questions were tested using the RDFLIB and the URLLIB

module in python. The results of the queries are shown in Table 6-2. The prefix for the ontology is denoted with "cs" in the queries in Table 6-2 for easy understanding and clarity.

Table 6-2: The SPARQL Results on the First Three Competency Questions with a Reflection on the Outcome. Source: Author

<table>
<tr><th>Question number (label is provided as [QSet. Question no])</th><th>Query</th><th>Results</th><th>Review</th></tr>
<tr><td>Q1.1

Which region is the highest reported number of species?</td><td>SELECT (DISTINCT ?o as COUNTY) (COUNT(?o) as ?No_of Sightings)

WHERE

{ ?s cs:FoundIn ?o . }

GROUP BY ?o

ORDER BY DESC(?No_Sightings)

Limit 3</td><td><table><tr><th>County</th><th>No_of Sightings</th></tr><tr><td>San Francisco</td><td>60312</td></tr><tr><td>Marin</td><td>56943</td></tr><tr><td>Napa</td><td>50660</td></tr></table></td><td>The query first selected the DISTINCT number of counties in the datasets based on the FoundIn relation defined in the ontology. It then counted the number of sightings in each County. The final operation ordered the different number of selected species ber County in descending order. (From highest to lowest). This shows that the datasets are made compatible with each other using the FoundIn relation in the ontology to group individual sightings from the different datasets into the different counties in California. The Result of Q1.1 is shown in Figure 6-6 with a proportional point symbol map showing the quantities of sightings per County</td></tr>
<tr><td>Q1.2

What are the land use and land cover characteristics of those regions?</td><td>SELECT (DISTINCT ?o as ?Land_Information)

{ ?s ?p "San Francisco" .

?s ? cs:FoundOn . }</td><td><table><tr><th>Land_Information</th></tr><tr><td>Forest</td></tr><tr><td>Grass</td></tr><tr><td>BareLand</td></tr><tr><td>Wetland</td></tr><tr><td>Transport</td></tr><tr><td>Pavement</td></tr></table></td><td>The query selected all information on San Francisco County in the RDF graph distinctively. It then selected the object properties that are mapped in the graph with different Land Use classes and landcover information for the selected county (San Francisco) based on the FoundOn Relation. This shows that the land use dataset, the Land cover and the species dataset are made compatible with each other.</td></tr>
</table>

Q1.3 Are there any risks of natural disaster reported in these areas?	SELECT (DISTINCT ?o as LandHazard) { ?s ?p "San Francisco" . ?s cs:HasRisk ?o . }	**LandHazard** Drought Landslides Earthquake	The query selected all information on San Francisco County in the data graph. It then selected the object properties that are mapped in the graph with different risks types using the relation HasRisk in the graph for San Francisco County only. The dataset presents information of the different land hazards in San Francisco. This shows how the different information on Hazards has been combined with other information of species. The combined information in the form of RDF can be used for making decisions.

Step 3: Land Cover validations on GeoWiki inputs

The validation of the Geowiki user input is based on the results obtained from the modelled dataset. Moreover, the modelled dataset is based on the semantics and the schema of the ontology. In the validation process, as shown in Figure 6-4, The result of Q1.1 is presented in **A**. This result serves as the input for Q1.2 and Q1.3. **B** and **C** in Figure 6-5 aim at giving a visual representation of the results of Q1.2 and Q1.3. Moreover, different species were modelled to achieve the final combined dataset. **D** aims at showing the different species used.

In the validation process, the ontology generates different latitudes and longitude based on the ***HasLong*** and ***HasLat*** relation defined in the ontology as shown in Figure 0-5 in the Appendix. The strategy deployed was plotting the longitude and latitude on a land use input. The result is shown in Figure 6-5.

Considering Figure 6-4, the **A** represents the overall county of San Francisco whiles **B** represents a zoomed in version of **A**. Most information on the modelled dataset appears to be in the open space class. This may be that most of the sightings information are on recreational parks and other open areas.

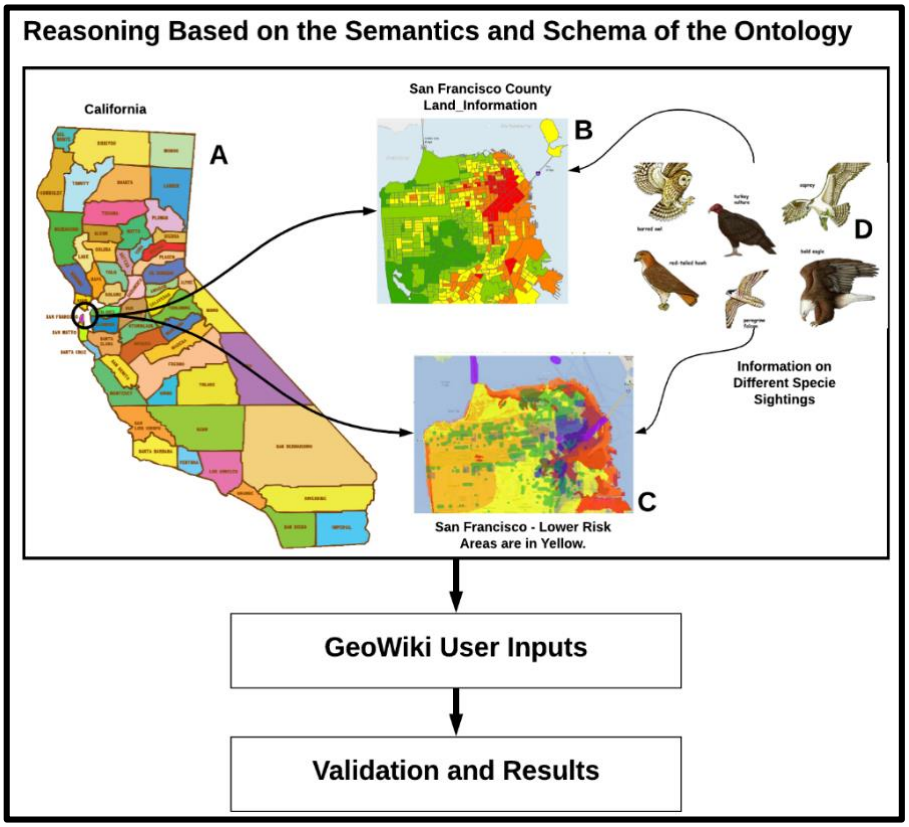

Figure 6-4: Overall Implementation Strategy for Set One of Use Case: Validating User Input from GeoWiki Based on the Semantics and Schema of the Citizen Science Ontology: Source: Adapted from (USGS, 2010)

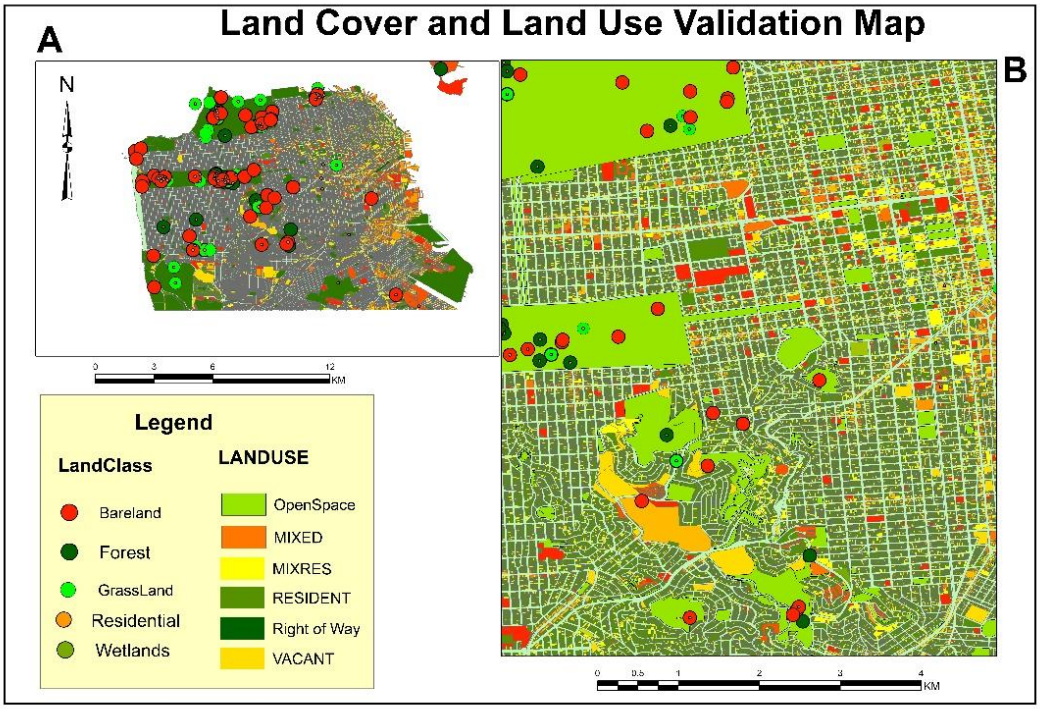

Figure 6-5: Different landcover information plotted on the different land use information to validate the potential use of the ontology: Source: Author

6.2.2. Semiotic Metric Suite

This section evaluates the designed ontology using some list of selected metrics based on the semiotic theory. The selected metrics checks for the Syntactic, Semantic and Pragmatic quality of the designed ontology and its proposed by Burton-Jones at el (2005). Table 6-3 shows a general overview of the selected metrics for assessing the quality of the ontology. It was adopted and modified from the semiotic framework proposed by (Burton-Jones et al., (2005), p. 6

Table 6-3: Overall Semiotic Quality Testing Strategy. Source: Adapted from (Burton-Jones at el., 2005)

Overall Metric	Main Metric Suite	Attributes	Description
Ontology Quality	Syntactic	Lawfulness	Correctness of syntax used compared to the OWL 2 Syntax
		Richness	Some features and syntax used compared to the total number of feature.
	Semantic	Interoperability	Meaningfulness of terms and concepts captured
		Consistency	Uniformity of the terms used in the ontology
		Clarity	Average number of word sense
	Pragmatic	Comprehensiveness	Number of classes and relation

Syntactic Quality

The syntactic quality describes the readability and conformity of the ontology. This is by comparing the structure and rules in the ontology to a proposed standard (OWL 2 syntax). It assessed the quality of the ontology using two sub-scores. These sub-scores are the lawfulness and the richness of the ontology.

Lawfulness (L): The lawfulness of the ontology determines how the designed ontology conform to the syntax in the OWL 2 syntax Amith & Tao, (2017). The lawfulness of the ontology is a score of the ontology based on the syntax used. The final Citizen Science ontology is implemented in the *OWL 2* syntax. Therefore, the lawfulness is a measure of the ratio of axioms used in the ontology that are in violation of the OWL 2 syntaxes proposed by *W3C* as compared to the total axioms created in the ontology. There are 36440 axioms currently in the Citizen Science ontology. The protégé editor incorporated with the *OWLAPI* plugin shows the total number of axioms in violation to the *OWL 2* syntax in the Ontology Matrix section. Figure 0-3 in the Appendix shows the ontology Metrix from the protégé environment. There was zero number of axioms in violation of the ontology from the OWL API plugin.

The protégé environment and the *HermiT 2* reasoner used in the design of the ontology checks for inconsistency in the ontology as well as concept formulations and statements that violate the *OWL 2* syntax. The editor provides means of solving such inconsistencies before the reason can efficiently operate. Therefore, all inconsistencies are resolved in the validation section. There are currently no violations of the ontology according to the OWL 2 syntax.

Richness (R): The richness of the ontology compares the number of features or ontological classes created in the ontology and the ones utilised by assigning different classes and data properties Amith & Tao, (2017). There are precisely 3850 classes in ontology as shown in Figure 0-3. Out of this number, **3588** classes are mapped with different relations and attributes. The SubClassOf relation was not considered in the relation count among the different classes. The computation of the richness score is shown in Table 6-4. The final

score value is **0.930016.** The value shows that about 90% of the generated classes are mapped with different relation both among classes and instances. Note **A** and **B** in Table 6-4 is obtained from the class, data property and object property matrixes in the protégé environment

Table 6-4: Calculating the Semantic Richness of the Ontology Based on the Semiotic Theory. Source: Author

Semantic Richness Score		
Number of Classes (A)		3858
Number of Classes Unused (B)		270
Number of classes Used (D)	(A-B)	3588
Richness Score (S)	(D/A)	0.930016
Percentage Score	S*100	93.00%

Semantic Quality

The semantic score of the ontology evaluates the ontology based on the terms and labels defined and used in the ontology. The terms in the ontology express the semantics of the concepts in the ontology which serve as the foundation for modelling different datasets. It assesses the ontology based on three sub-modules as shown in Table 6-3. These three sub-scores are the Interoperability, Consistency and Clarity.

Interoperability: In determining the interoperability score of the ontology, it was rated by computing the percentage of randomly selected terms in the ontology with word sense Amith & Tao, (2017). Twenty (20) words (Classes) were selected at random using the *NUMPY* and *RDFLIB* modules in python. These words were checked in another existing semantic data source (Wordnet) using the *RDFLIB* and the *NTLK* module for natural language processing. Each word is checked for its definition on the WordNet platform. Each definition is compared to the proposed semantics in the ontology. The results of the twenty selected words are shown in Table 6-5. It was realised that; all twenty randomly selected words to have similar natural meaning as compared to the meaning and how they are used in the ontology. However, some of the concepts in the ontology does not have a direct definition of the wordnet platform which results in errors during the search. Therefore, the process became irritative till all randomly selected words had meanings on the wordnet platform. The interoperability is calculated by using the ration the number of terms with word sense in the randomly selected words to the total number of randomly selected terms.

Table 6-5: Comparison of Words meaning (Semantics) in the Ontology to the WordNet Platform semantics: Source Author

Number	Word	Semantics in the ontology	Wordnet Meaning	Word Sense
1	Mangrove	Various types of salt-tolerant plant species (trees or shrubs) that occur in intertidal zones of tropical and subtropical sheltered coastlines. The term is applied to both the individual plant and the broader ecosystem	A tropical tree or shrub bearing fruit that germinates while still on the tree and having numerous prop roots that eventually form an impenetrable mass and are important in land building	1

2	Adaptation	Adjustment in natural or human systems to a new or changing environment.	A written work (as a novel) that has been recast in a new form	3
3	Biome	The most significant unit of ecological classification that is convenient to recognise below the entire globe	A major biotic community characterised by the dominant forms of plant life and the prevailing climate	1
4	Desert	Degradation of land in arid, semi-arid and dry sub-humid areas, resulting from various factors, including climatic variations and human activities	Arid land with little or no vegetation	1
5	Community	A community of plants and animals characterised by a typical assemblage of species and their abundances	A group of people living in a particular local area	6

Consistency: The Consistency module checked for the total number of duplications in the ontology. It is computed by finding the total percentage of terms that are consistently the same across the ontology (duplications) Amith & Tao, (2017). This was computed by selecting all the number of classes in ontology using the *RDFLIB* module and SPARQL for sorting the classes in a text file. The number of classes was selected by counting and grouping the distinct words occurring in the text file. 30% of the words have double occurrences in the ontology. These duplications were as a result of the groupings of concept into both domain-specific concepts and generic concepts.

Clarity: The clarity module checked for how unambiguous the terms in the ontology are used. This was done by randomly selecting a list of 20 words (in the ontology and comparing these randomly selected words with the meaning of the words on the WordNet platform Amith & Tao, (2017). It reveals that out of the twenty words, more than 60% of the words had more than one meaning. These words have different meaning and are mostly based on how each term is constructed in a sentence. Therefore, it is clear that defining classes in a diverse domain should have more clarification on the intended purpose of the classes when created and should be made clear to the user. Word sense result is displayed in Table 6-5 and others in Table 0-1 in the Appendix

Pragmatic Quality

The pragmatic qualities in this section only consider the comprehensiveness of the ontology. This is because the other modules in the semiotic framework such as the relevance are being considered in the "use case sections".

Comprehensiveness: This indicates the sizes of the ontology. Larger ontologies are likely to capture the most concept in the domain they represent. Therefore, the larger the size of the ontology, the more likely the ontology expresse most knowledge in the domain of Citizen Science. The size of the ontology is defined by the total number of classes and the total number of relations in the ontology. From Figure 0-3 in the Appendix, the total number of classes in ontology stands at 32440 with more than 423 relations. Therefore, the size indicates that the ontology expresse several concepts in the domain of Citizen Science.

6.3. Deployment and Maintenance

The act of deploying a tool or a system is to release the system with adequate information for users in the community and purpose it was designed. This section describes a system for releasing the designed ontology for the Citizen Science community. A peer to peer review and an open source system for releasing the ontology to the Citizen Science community is considered in this section.

Publishing and usage: The intended users of the Citizen Science ontology are the researchers and data scientists in the Citizen Science community. The final product of the ontology is the OWL/XML file. The file is attached to the Karma data integration tool for easy usage. The combine ontology, use cases datasets and description are uploaded to the GitHub repository. The repository is organised and save with the name CitizenScienceOntology for easy identification and modification by members. The documentation on the ontology is added to the wiki section of the Git repository. A screenshot is displayed in Figure 0-3 in the Appendix. This is a link to the git repository[29]

Maintenance: The maintenance of the Citizen Science ontology is set to be done using semantic peer to peer (P2P) systems for sharing and updates on the GitHub platform. This is achieved using the git repository on the GitHub platform. Semantic P2P systems are open source platforms with a connected network of groups of experts both in technical and domain-wise (Staab et al., 2006). These experts are willing to develop and use applications and systems to contribute to the development of technology. The platform provides a means for controlling different platforms of the Citizen Science ontology.

[29] https://github.com/CitizenScienceOntology/Ontology

7. DISCUSSION

The design and prototyping of an ontology for integrating Citizen Science datasets proves not only that a well-structured and organised vocabulary of concepts about a domain (Citizen Science) can support data integration but can also improve rectifying different datasets to ensure compatibility and ease sharing of the datasets on the semantic web. This section aims at giving an overview of the logic and inferences that can be obtained from the design of the Citizen Science ontology. It is structured to discuss the overall work done in this thesis by taking into consideration the research objectives and research questions. It presents the criteria reviewed for selecting a sound methodology, an overview of the assumptions made during the design and the results obtained from the quality testing both from the use case perspective and the semiotic metric suit testing.

7.1. Criteria and Methodology Selection

There are several criteria according to literature for the design of ontologies. Most of these criteria evolved based on the domain and purpose for the envisaged ontology (Gruber, 1995). This section gives some general discussions on the type of choices that informed the selection of the methodology for the design of the Citizen Science ontology.

Several reports have shown that the number of criteria for building ontologies has grown significantly. It is possible to infer that the increases in the number of criteria are likely due to the aim of providing more capabilities of both ontology methodologies and ontology design process. The selection of the criteria is made to ease the difficulties in the design process and for an efficient ontology quality. This project selected the most relevant criteria based on their frequent occurrences in literature as well as the range of domain that they are applicable. What is surprising is that, after considering several methodologies, not a single method considered all the selected criteria. Moreover, few methodologies considered potential means of encoding geospatial information in the ontology design process. This confirms the consideration of the purpose of ontology before selecting a potential ontology methodology.

The main challenge that affected the selection of the criteria was on the impact of each criterion on the designed methodology both technically and technologically. An example is the formularization criterion which has several proposed formularization strategies from the different methodologies. In the selection of a methodology, how each methodology formularies the intended domain concepts in the ontology was considered. Example; the W3C GEO Vocabulary formalises the domain concepts with an emphasis mostly on geographic concepts. Such a methodology, when selected for a diverse domain of citizen science will not only affect the concept capturing efficiency but will turn to overlook several vital concepts which have no spatial information contained in them. Technologically, few ontology editing platforms have the capability of directly acquiring different knowledge from different sources. This limitation of ontology editing tools in the ontology engineering hiders the efficiency of the envisaged of the ontology. Therefore, a methodology like the DOGMA which proposes the use of the DOGMA Studio Workbench which has few knowledge acquisition capabilities will only limit potential concepts needed to formularies a diverse domain. (Technological limitations).

The results obtained from the criteria selection further support the idea of Liu et al., (2011) and Lozano-Tello & Gómez-Pérez, (2004) which proposed considering the domain and purpose the ontology is going

to serve. In general, Citizen Science can be considered as an approach to science which covers a broader range of different scope of sciences. Therefore, in the selection of the methodology for developing ontologies for such a diverse domain, considerations on applicability, purpose, domain usage and the number of occurrences is required. The selected criteria range from reuse capabilities (which present the possibility to use different existing ontological concepts) to geospatial capabilities (ability to encode geospatial information) as discussed in Section 3.3. These criteria were considered as good indicators for a potential methodology for designing quality ontology for the Citizen Science domain.

Therefore, in selecting a list of criteria for designing ontologies, there should be considerations on the domain, the type of ontology to be designed and the purpose of the envisaged ontology. It may in-turn enhance the performance of the ontology quality as well as the easiness in building the ontology from scratch.

7.2. Design and Implementation

The design of the Citizen Science ontology used the IEEE standard for software development life cycle fused with the Generic Ontology Development Framework. The merging of the two frameworks gave the structure of the design a solid foundation to inculcate all necessary and potential information in the design process. The final designed ontology comprises a list of concepts expressed in the OWL 2 syntax for the domain of Citizen Science. It is stored internally in the developing environment of protégé with the HermiT_2 reasoner as an implementation strategy. This section aims to give a general discussion of the design of the ontology.

The management activities gave the design a stable structure due to its comprehensive coverage and overall representation of the domain of Citizen Science. It emphasised on the selection of Upper-Level concepts which were mostly based on the review of the domain of Citizen Science in Chapter Two. The relevant sections of Chapter Two served as the Upper-Level concepts in the design process. The selection of the Upper-Level concepts supported wider interoperability among the different datasets semantically. Therefore, defining an unambiguous Upper-Level concept in an ontology design process can be seen as a means of enhancing the formulation and definition of the different concepts (Sub-classes under each Upper-Level concept). The management activities served as the foundation for the development stage.

The development stage includes document specification and data source determination sections. It discusses the general criteria for selecting potential information adequate for providing knowledge to be used for the formularization of the ontology. The different data source obtained had different data formats. This led to the three different Knowledge acquisition strategies used in the knowledge acquisition stage. In effect, several datasets formats can exist which may require different knowledge acquisition strategies other than the ones considered in this thesis work. Therefore a potential area of research may be an investigation into defining a generalised strategy that can process all the different datasets into the ontology in one instance.
The design process of the ontology had several stages with most stages iterative due to the two fused frameworks. The iterative stages required several pre-processing of the different information to yield a concrete conclusion before moving to the next stage. Moreover, the output of the current stage in the design process served as the input for the preceding stage. Most of the iterations occurred in the knowledge acquisition stage. This is due to lack of common standard and pattern among the different datasets converted to OWL Ontologies.

Due to overlaps of concepts in the domain of Citizen Science, several duplications of concepts occurred during the data acquisition stage. These duplications were manually removed to avoid inconsistencies and to ensure uniformity in accordance with the OWL-2 syntax. The compiled OWL ontologies were cross-checked to provide an automatic means of defining class-subclass hierarchy in the ontologies during the ontology formularization. It was quite clear that the semantic association across the different generated OWL ontologies were not the same. Therefore, the class-subclass hierarchy was defined manually. The manual definition provided logical flow and consistency in the ontology formularisation stage. Class and subclass relations were developed based on the chosen semantic structure. In effect, defining a particular ontology semantic structure in ontology design can be considered as an efficient means of ensuring logical consistency in the ontology design process.

The user requirement analysis at the design stage only considered the ontology not to be a stand-alone tool but a tool to be used in other development environments. Therefore, the user requirement considers only proper documentation and clarity of concepts in the design of the ontology. The documentation in the ontology was done by annotating different classes in the ontology with the semantics each class portrays.

Furthermore, the design process was quite bulky. Therefore, most of the classes and relations created in the ontology were not mentioned in the reports. The different classes and relation, as well as their capabilities, can be cross-checked from the documentation folder on the Git repository. In this way, the different sections required to integrate different datasets can be analysed and use accordingly. Since the ontology is still in the developmental stage; several strategies were deployed to provide possible means of extending the concepts defined in the ontology.

7.3. Quality Testing

The quality testing was performed on the ontology and the output of the ontology after it was used as a surrogate for modelling different datasets. The testing of the modelled datasets was done to ensure compatibility among the different datasets used in the modelling stage. This section considered the result of the quality testing both the use case testing and the semiotic metric suite.

Use Case Testing

The use case testing tested for the different parts of the information from the different datasets contained in the combined dataset using the competency questions translated into SPARQL queries. This section aims to discuss the first three competency questions and their results to show the added value of the ontology to the data integration paradigm.

From the SPARQL results, it seems possible that the ontology can be used to perform several activities regarding data management and data integration. These activities include some traditional GIS functionalities as well as well-structured database operations. Question 1 serves as the confirmation for some of these capabilities. The combined dataset consists of an interconnected RDF graph (dataset). In question 1, the ontology was used to organise the different datasets using the ***SubClassOf*** relation in the ontology. The different datasets on species were combined in a hierarchy under the class **Species** in the ontology. This means all the different species were considered as instances ***Type_Of*** the **Species** class proposed in the ontology. The organisation of the datasets according to the semantic and schema of the ontology followed the axioms **OWL** ***SubClassOf*** **Species**, **Dragonfly** ***SubClassOf*** **Species** and many others. Results of question 1 is in agreement with the qualities propose by Masolo & Borgo, (2005). Using the ***FoundIn*** relation defined in the ontology, different species were grouped in the combined dataset into

the different counties in California. The SPARQL query then selected the County with the highest number of species from the combined dataset (San Francisco). The output of Query 1 served as the input for query 2. That means the "Where" clause in query 2 selected all the different land use and land cover information for only San Francisco County.

The query aims at testing for land use and land cover information in the combined datasets. The question analysed the different land use and land cover classes contained in the combined datasets and selected only classes which present land use and land cover information in San Francisco County.

Therefore, the different land use and land cover information were made compatible with the species dataset with the combined information (Added Value). In effect, query two tested for the added value of the ontology in the data integration paradigm. From Figure 0-6 in the Appendix, the query operated on only San Fransico county for the different land cover information.

The output of query one was also used in query 3 to check for different natural hazards that are predominant in San Francisco County for the areas of the reported species. The objective is to test for information on natural hazards from the combined dataset. Since datasets on different natural hazards were used in the modelling process. The results prove that the different information on natural hazards can be obtained in the selected county.

Question 3 also takes the output of question 1 as input and computes the risk-averse areas in San Francisco County. Therefore, query 2 and 3 can be said to have an added value of the designed ontology as compared to question one. However, the result of question one serves as the input of question 2 and 3. A detail connecting strategies for connecting different datasets based on the ontological classes are shown in the Appendix Table 0-3. The results when combined was for validating land classes which include land use and land cover information. The results shows that when different species are carefully modelled in an ontology, the resulting information can be used for validating land information by givin the actural landclass at that vicinity. In conclusion, adequate information on each dataset can be found in the combined dataset. Which can be used to yeild different information.

Semiotic Metric Suite Testing

As the meaning of semiotic implies, the quality testing with the semiotics, tested mostly for the signs and symbols used in the ontology. The semiotic metric suite test tested for both the intrinsic and extrinsic quality of the designed ontology. The test was organised under three quality indicators (modules) proposed by the semiotic framework for ontology evaluation (Burton-Jones et al., 2005). The three quality indicators discussed in section 6.3.2 and they are the syntactic, the semantic and the pragmatic quality of the designed ontology. This section aims to discuss the results of the semiotic metric suite testing.

The syntactic quality scored the ontology based on the **lawfulness** and the **richness** of the ontology. The lawfulness of the ontology as a quality indicator was assessed by comparing all the syntax of axioms in the ontology to the OWL 2. Syntax. Moreover, the ontology formularisation was done concurrently with the HemiT 2 Reasoner running. It was quite evident that none of the axioms was in violation of the OWL 2 syntax since the ontology was carefully designed and implemented with The HermiT 2 reasoner. This confirms the capabilities of the HermiT 2 reasoner proposed by Glimm et al., (2014).

The richness quality indicator scored the ontology based on the number of ontological classes generated in relation to the used ones among these classes. After the comparison with the OWL API in the protégé environment, the ontology scored more than 90%. This implies the designed ontology can be considered to have an appropriate intrinsic quality by having a higher syntactic score.

The semantic quality tested for the quality of the terms and labels used in the ontology. It scores the ontology on three subscores. These sub-scores are the consistency, clarity and interoperability. Each score was computed based on the requirement proposed by the semiotic framework. The results of the interoperability score were generated based on the semantics of concepts in the ontology compared to the semantics of the same concepts on the WordNet platform. However, the assumption made was, all the meaning of words on the Wordnet platform was considered to be accurate and expresses the knowledge through the definition of the word (Word Meaning). Moreover, it was seen that; a single word might have more than one meaning depending on its usage in a part of speech. Therefore, direct comparison of the meaning of the words was avoided. The meaning of each word was chosen based on a conceptualisation that reflects the domain of Citizen Science. The results can be seen in Table 0-1 in the Aappedix. It can be concluded that some words have no single meaning unless they are in action. Therefore, expressing such words in the ontology can results in ambiguity if not defined. This confirms the results of Amith & Tao, (2017) during it evaluation using semiotics

The consistency score evaluated the ontology based on the number of duplications in the ontology. An ontology is a graph of interconnected concepts that describe a domain. Therefore, duplications are mostly hard to avoid. The designed ontology has duplications in the concepts expression stage due to the grouping of concepts into domain specific and generic. Therefore, in general, the score of the ontology on consistency is reasonably average for the ontology since few duplications were determined.

The clarity score evaluated the ontology based on the expressiveness and unambiguous in the ontology.

The WordNet platform used as a knowledge-based showed that most words have word sense of an average of 2. The strategy deployed to address the issues with clarity in the ontology is by connecting words in the form of phrases to define a class. Examples are '**Concept Scheme', EarthquakeStation** among others. Using phrases as concepts in the ontology can be viewed as means of clearing certain ambiguity in the ontology.

The pragmatic score considered only the comprehensiveness of the ontology. However, from the concept of ontologies, comprehensiveness cannot be entirely determined due to the unlimited number of concepts in a domain. However, it can only be assumed by considering the size of the ontology.

7.4. Deployment and Maintenance

The ontology deployment considered the open source system of knowledge sharing. In effect, it can be considered as an effective means of releasing specific knowledge about any phenomenon to mass media who might be interested in the product under consideration. The GitHub platform has more than 27 million subscribers. These subscribers are willing to edit, use and maintain resources and projects hosted on the platform with the Citizen Science ontology, not an exception.

8. CONCLUSIONS AND RECOMMENDATIONS

8.1. Conclusions

This thesis collected and reviewed information on the established concept and practice of citizen science and provided information on aspects relevant for developing applications using Citizen Science data, data sources and data characteristics. It reviewed ontologies and ontology design by describing a process of selecting a suitable methodology for designing ontologies for a diverse domain. The selection of the ontologies was based on frequently used criteria that appear in literature. Several different approaches towards creating new ontologies from scratch are outlined, their characteristics identified, and their suitability for applying them to the domain based on most frequently used criteria and principles in ontology design is evaluated. Moreover, the Generic Ontology Development Framework turns out to be the best-fitting approach in the context of Citizen Science ontology. This framework is used in this work to presents a procedure for generating OWL ontologies from different data formats and characteristics based on a selected methodology. The thesis adapted a semantic structure from the framework for organising classes and properties in developing the Citizen Science ontology. The Citizen Science ontology is built with OWL 2 syntax. Moreover, it appears to be a comprehensive and detailed conceptualisation of some critical aspect of the Citizen Science domain. The evaluation of the designed ontology considered two approaches. The Data-driven approach "use cases" and the intrinsic and extrinsic evaluation approach through the use of a semiotic metric suite. During the evaluation stage, a data-driven approach was adopted by selecting different use cases to test the quality of the ontology. The Citizen Science ontology proved to be competent since all the generated queries resulted in the required information. Selected problems regarding data interoperability and reuse of datasets in Citizen Science can be modelled with the ontology due to its capability to serve as a surrogate for modelling different dataset to one consistent and comprehended data. The ontology provides a well-structured and adequate information needed to integrate different dataset in Citizen Science. However, the process of modelling the different datasets to make them interoperable was not fully automated due to challenges in semantics conversion which includes natural language processing. (Transforming different datasets to accept match classes, properties and datatypes and during the modelling stage). Semantic data integration by merging different datasets using an ontology as a surrogate as well as enabling semantic reasoning and querying is an essential component in semantic web technology for Citizen Science community. This thesis has mapped potential information using the OWL 2 syntax to characterise and express knowledge in the domain of Citizen Science. Different datasets can be made compatible with each other and be shared, reused and extended for different applications when modelled the designed Citizen Science ontology.

8.2. Answers to Research Questions

Objective One

1: What are the criteria for selecting a methodology for the design of the ontology?

There are several criteria for selecting a sound methodology for building ontologies. However, some appear to be more useful due to its frequent usage and appearance in literature. Examples include Reuse Capabilities, Consistency, Geospatial Capability, Formularization, Completeness, Interoperability, Co-creation and Modularization. Details are expressed in Section 3.4

2: What are the key components and the principal requirements for ontology design?

The requirement for designing ontologies are quite extensive depending on the domain and purpose of developing the ontologies. From the design and the implementation in chapter five, the key and most important component of ontology design is a clear understanding of the domain the ontology is intended for. However, other components of equal importance include data and information availability, defining a

clear scope, competency, granularity, clarity in the domain details of each component can be found in Section 5.3.1.

3: How can the principles behind ontologies be applied to concepts in the Citizen Science domain?
Applying ontological principles to concepts in Citizen Science can follow any normal procedure for capturing knowledge for ontology in a domain. However, deferent processing needs to be done to avoid problems in processing natural language semantics, where different words and expressions are used to imply different meanings and conditions. An adequate and potential means of defining classes should always avoid inconsistencies in the domain of Citizen Science. This is by defining a clear semantic structure for organising classes and relations in the ontology details on the semantic structure can be found in Section 5.3.2.

Objective Two

1: What are the user requirements for the Citizen Science ontology?
The use of the ontology at this stage was not considered as a standalone tool for data integration. Therefore, user requirements for data integration considered the selected tool for using the ontology. The characteristics of the selected tool for data integration using the ontology as a surrogate must be based on the criteria reviewed in Section 6.1.1 Table 6-3. The requirements include technical support, search capabilities, analyser[30], simplicity in usage, capabilities, the efficiency of the tool, proper documentation and ability to both debug and modify the tool as well as the steps in modelling the datasets during the data integration process. However, the final ontology should have maximum documentation and precise definition of classes in the design stage. These two criteria can be considered as a potential requirement specifically for the designed ontology.

2: What are the criteria for defining ontological classes in the Citizen Science ontology?
The criteria used for defining classes followed the semantic structure proposed in the Generic Ontology Development Framework. The structure defines explicitly what constitutes a class and a relation in the ontology. Details on the class definition and relationship definition can be obtained in Section 5.3.2. Under the semantic structure and ontology formularization section.

3: How will the relationships between classes in Citizen Science be established?
The relationships among classes were defined manually due to the inconsistent semantic association in the acquired knowledge (concepts and classes). Semantic association of the different sections and different knowledge contained in the generated OWL ontologies across all resources were not the same. Therefore, the class-subclass hierarchy and the relationships between different classes and entities were defined manually. The manual strategy is provided more logical, coherent and consistent semantic association among concepts. However, some relations occurred naturally due to the natural groupings in the datasets. The relationships among classes and individuals are realised based on the natural groupings that exist in the datasets. New classes evolved as the different datasets communicate different form of knowledge which can be considered as concept or class in the ontology

4: What are the requirements for implementing the Citizen Science ontology?

[30] The ability of the tool to examine in details a given data to help determine common patterns and relations that exist in the dataset.

The implementation of the ontology considered several requirements. These requirements include the selection of an editor and a reasoner with a list of qualities stated in Section 5.4.1. Example include

1. A lexical and syntactic analyser to guarantee the absence of lexical and syntactic errors.
2. An editor for adding, modifying, and removing definitions.
3. A browser for inspecting the library of ontologies and their definitions.
4. Evaluators for detecting incompleteness, inconsistencies, and redundant knowledge.

Objective Three

1: What are the strategies for testing the quality of the Citizen Science ontology?
There are no accepted standard criteria for testing the qualities of ontologies. However, there are several proposed strategies potential for evaluating ontologies. This research considered two of the criteria. These evaluation criteria are discussed in Chapter 6. It includes use cases and semiotic metric suite testing.

2: What are the quality criteria to be used for the metric suite testing?
The quality criteria considered in the metric suite testing were based on the semiotic theory. The selected criteria are the Syntactic Quality criteria, Semantic quality criteria and Pragmatic quality criteria. Details are discussed in Section 6.3. The selection of the quality indicators was to reflect and test both the intrinsic and extrinsic qualities of the designed ontology. The results from these quality indicators showed how the ontology can be checked and assessed to make them compatible to other domains for reuse.

3: What are the strategies for integrating the Citizen Science ontology into the mainstream ontologies?
This is not considered Yet: From the design and reuse stage in Chapter Five, integrating the current ontology to the main stream ontology will require ontology alignment and ontology mapping. Where common vocabularies are mapped to the main stream ontologies. When the URI of resources in the ontology are mapped during the reuse section with the prefix of the reused ontology class, this different URIs can be assessed and used as the matching strategy with other.

Objective four

1: How will the developed Citizen Science ontology be published?
The ontology is Published in a Git repository on the GitHub platform as an open source tool for easy editing and usage.

2: What are the strategies for maintaining the developed Citizen Science ontology?
The maintenance of the ontology is presumed to be done on the git repository by interested members in Citizen Science out of all the 27 million subscribers in GitHub Community. Therefore, using a peer to peer network system for review on the GitHub repository is the key maintanance strategy. The peer to peer system on GitHub is a structured PNP system that achieves its usage and reviews based on a dynamic maintenance and update system. It is envisaged that every citizen scientist who uses an open source application when searching for application on the GitHub platform will choose a keyword which in clude Citizen Science. Such search with the Key term citizen science on the GitHub platform will have the ontology in the drop-down list for easy selection.

8.3. Recommendations and Future Work

A future point of interest in this research are A). the investigation into a possible means of automating the modelling of the datasets using the ontology. That is by designing a modelling tool the can automatically learn from previous modelling strategies to integrate different datasets based on the Citizen Science ontology semantic and schema. B). An automatic means of inculcating different extraction processes for different data formats to enhance efficient knowledge extraction from datasets and other sources. C). Exploring more details on the geospatial capabilities of the Citizen Science ontology using GeoSPARQL technology. D). Developing the ontology further and validation by the CS community and mapping with other ontologies.

LIST OF REFERENCES

Alam, M. D., & Gühl, U. F. (2016). Project Phases. In *Project- Management in Practice* (First, pp. 55–121). GmbH Germany: Springer-Verlag. https://doi.org/10.1007/978-3-662-52944-7

Amith, M., & Tao, C. (2017). Modulated evaluation metrics for drug-based ontologies. *Journal of Biomedical Semantics*, *8*(1), 1–8. https://doi.org/10.1186/s13326-017-0124-2

Auer, S. (2006). RapidOWL - An Agile Knowledge Engineering Methodology. *Perspectives of Systems Informatics, 6th International Andrei Ershov Memorial Conference, PSI 2006*, 424–430. https://doi.org/10.1007/978-3-540-70881-0_36

Avraham, U., Hasson, A., & Matti, R. (2013). Introduction to logic and set theory.

AWS. (2010). Home | ZomBee Watch. Retrieved October 16, 2017, from https://www.zombeewatch.org/

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, *37*(1), 61–81. https://doi.org/10.1007/s10115-012-0571-0

Barbosa, L., Pham, K., Silva, C., Vieira, M. R., & Freire, J. (2014). Structured Open Urban Data: Understanding the Landscape. *Big Data*, *2*(3), 144–154. https://doi.org/10.1089/big.2014.0020

BBC. (2017). BBC - Ontologies - Wildlife Ontology. Retrieved December 19, 2017, from https://www.bbc.co.uk/ontologies/wo

Bhattacharjee, Y. (2005). Ornithology. Citizen scientists supplement work of Cornell researchers. *Science (New York, N.Y.)*, *308*(5727), 1402–1403. https://doi.org/10.1126/science.308.5727.1402

Bleumers, L., Jacobs, A., Sulmon, N., Verstraete, M., Van Gils, M., Ongenae, F., … De Zutter, S. (2011). Towards Ontology Co-creation in Institutionalized Care Settings. *Proceedings of the 5th International ICST Conference on Pervasive Computing Technologies for Healthcare*, 559–562. https://doi.org/10.4108/icst.pervasivehealth.2011.246110

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bowser, A., Wiggins, A., & Stevenson, R. (2013). Data policies for public participation in scientific research: A primer, 1–13. Retrieved from http://www.dataone.org/sites/all/documents/DataPolicyGuide.pdf

Boyce, S., & Pahl, C. (2007). Developing domain ontologies for course content. *Educational Technology and Society*, *10*(3), 275–288. https://doi.org/10.1007/s10791-006-9018-0

Brost, R. C., McLendon III, W. C., Parekh, O., Rintoul, M. D., Strip, D. R., & Woodbridge, D. M. (2014). A Computational Framework for Ontologically Storing and Analyzing Very Large Overhead Image Sets. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 1–10. https://doi.org/10.1145/2676536.2676537

Bryan, H. J., Kayri, H.-Y., & Ballard, J. S. (2017). Budburst Project. Retrieved October 24, 2017, from http://budburst.org/aboutus

Buccella, A., Cechich, A., Gendarmi, D., Lanubile, F., Semeraro, G., & Colagrossi, A. (2011). Building a global normalized ontology for integrating geographic data sources. *Computers and Geosciences*, *37*(7), 893–916. https://doi.org/10.1016/j.cageo.2011.02.022

BudGuide. (2013). Welcome to BugGuide.Net! - BugGuide.Net. Retrieved October 24, 2017, from https://bugguide.net/node/view/15740

Burgess, H. K., DeBey, L. B., Froehlich, H. E., Schmidt, N., Theobald, E. J., Ettinger, A. K., … Parrish, J. K. (2017). The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, *208*, 113–120. https://doi.org/10.1016/j.biocon.2016.05.014

Burton-Jones, A., Storey, V. C., Sugumaran, V., & Ahluwalia, P. (2005). A semiotic metrics suite for assessing the quality of ontologies. *Data and Knowledge Engineering*, *55*(1), 84–102. https://doi.org/10.1016/j.datak.2004.11.010

Casellas, N. (2011). *Legal Ontology Engineering*. *Media*. https://doi.org/10.1007/978-94-007-1497-7

Chenguang, L. E. E., Tdvrap, T., Sun, A., Fri, R., & Fri, T. (2015). Passengers â€™ information Flight details Contact / Billing information Price breakdown, *300*(917), 8–9.

Coffey, P. (1938). *Ontology or the Theory of Being. Psychology* (Vol. 19).

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information : The Nature and Motivation of Produsers. *International Journal of Spatial Data Infrastructures Research*, *4*(4), 332–358. https://doi.org/10.2902/1725-0463.2009.04.art16

Corsar, D., & Sleeman, D. (2008). Developing Knowledge-Based Systems using the Semantic Web. *BCS Int. Acad. Conf.*, 29–40. Retrieved from http://www.researchgate.net/publication/220847903_Developing_Knowledge-Based_Systems_using_the_Semantic_Web/file/32bfe511aa72c52876.pdf

COST. (2011). COST | The use of geographic information systems in Climatology and Meteorology. Retrieved August 17, 2017, from http://www.cost.eu/COST_Actions/essem/719

COST. (2013). COST | Climate Change Manipulation Experiments in Terrestrial Ecosystems - Networking and Outreach (ClimMani). Retrieved August 20, 2017, from http://www.cost.eu/COST_Actions/essem/ES1308

Da Silva, C. F., Médini, L., Ghafour, S. A., Hoffmann, P., Ghodous, P., & Lima, C. (2006). Semantic interoperability of heterogeneous semantic resources. *Electronic Notes in Theoretical Computer Science*, *150*(2), 71–85. https://doi.org/10.1016/j.entcs.2005.11.035

De Nicola, A., & Missikoff, M. (2016). A lightweight methodology for rapid ontology engineering. *Communications of the ACM*, *59*(3), 79–86. https://doi.org/10.1145/2818359

Ding, Y., Lonsdale, D., Embley, D. W., Hepp, M., & Xu, L. (2007). Generating ontologies via language components and ontology reuse. *Natural Language Processing and Information Systems, Proceedings*, *4592*, 131–142. https://doi.org/10.1.1.119.9828

Dröes, R. M., Bengtsson, J. E., Meiland, F., Haaker, T., Moelaert, F., Mulvenna, M., & Nugent, C. (2009). *Final Evaluation Report*. Brussels. Retrieved from www.cost.eu/COST_Actions/essem/719

Egenhofer, M. J., & Egenhofer, M. J. (1991). Point-Set Topological Spatial Relations. *The International Journal for Geographical Information Systems*, *5*(2), 161–174.

Eitzel, M. V, Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., … Jiang, Q. (2017). Citizen Science Terminology Matters: Exploring Key Terms. *Citizen Science: Theory and Practice*, *2*(1), 1. https://doi.org/10.5334/cstp.96

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching. ACMComputing Classification (1998): H.3, H.4, I.2, F.4* (1st ed.). New York: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49612-0

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *AAAI-97 Spring Symposium Series*, *SS-97-06*, 33–40. https://doi.org/10.1109/AXMEDIS.2007.19

Figueiredo, S., Cuccillato, E., Schade, S., & Guimaraes Pereira, A. (2016). *Citizen Engagement in Science and Policy-Making*. https://doi.org/10.2788/40563

Flowerdew, R. (1991). Spatial Data Integration. *Geographic Information Systems and Science*.

Follett, R., & Strezov, V. (2015). An analysis of citizen science based research: Usage and publication patterns. *PLoS ONE*, *10*(11), 1–14. https://doi.org/10.1371/journal.pone.0143687

Fonseca, F., & Martin, J. (2007). Learning The Differences Between Ontologies and Conceptual Schemas Through Ontology-Driven Information Systems. *Journal of the Association for Information Systems*, *8*(2), 129–142. https://doi.org/Article

Frechtling, J. (2002). An Overview of Quantitative and Qualitative Data Collection Methods. *The 2002 User-Friendly Handbook for Project Evaluation*, 43–62. Retrieved from http://www.nsf.gov/pubs/2002/nsf02057/nsf02057.pdf

Geoghegan, H., Dyke, A., Pateman, R., West, S., & Everett, G. (2016). Understanding Motivations for Citizen Science. Final Report on behalf of the UK Environmental Observation Framework (UKEOF), (May), 124. Retrieved from http://www.ukeof.org.uk/resources/citizen-science-resources/MotivationsforCSREPORTFINALMay2016.pdf

George, D. (2005). Understanding structural and semantic heterogeneity in the context of database schema integration. *Journal of the Department of Computing, UCLAN*, *4*, 29–44. Retrieved from http://www.thewebtrain.co.uk/portfolio/Resources/Heterogeneity_DoC_Conf_ 4_May_05.pdf

Gil, R., & Martin-Bautista, M. J. (2014). SMOL: A systemic methodology for ontology learning from

heterogeneous sources. *Journal of Intelligent Information Systems*, *42*(3), 415–455. https://doi.org/10.1007/s10844-013-0296-x

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., & Wang, Z. (2014). HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning*, *53*(3), 245–269. https://doi.org/10.1007/s10817-014-9305-1

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Graham, L. J., Haines-Young, R. H., & Field, R. (2015). Using citizen science data for conservation planning: Methods for quality control and downscaling for use in stochastic patch occupancy modelling. *Biological Conservation*, *192*, 65–73. https://doi.org/10.1016/j.biocon.2015.09.002

Grau, B. C. (2010). Modularity and Web Ontologies. *Springer*, 198–208.

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. https://doi.org/10.1006/ijhc.1995.1081

Guarino, N., & Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases. Knowledge Building and Knowledge Sharing*, *1*(9), 25–32. https://doi.org/10.1006/ijhc.1995.1066

Hadj, T., Ali, M., Aouicha, B., Hamadou, M., & Abdelmaji, B. (2014). Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence*, *36*, 238–261. https://doi.org/10.1016/j.engappai.2014.07.015

Hadzic, M., Wongthongtham, P., Dillon, T., & Chang, E. (2009). *Ontology-Based Multi-Agent Systems*. *Studies in Computational Intelligence* (Vol. 219). https://doi.org/10.1007/978-3-642-01904-3

Hand, E. (2010). Volunteer army catches interstellar dust grains. *Nature*, *466*(August), 685–687. https://doi.org/10.1038/news.2010.106

Hanski, I. (2016). A Practical Model of Metapopulation Dynamics Author (s): Ilkka Hanski Published by : British Ecological Society Stable URL : http://www.jstor.org/stable/5591 REFERENCES Linked references are available on JSTOR for this article : You may need to log in t. *Animal Ecology*, *63*(1), 151–162.

Hendler, J. (2014). Data Integration for Heterogenous Datasets. *Big Data*, *2*(4), 205–215. https://doi.org/10.1089/big.2014.0068

Hoedjes, J. C. B. (2014). *Public Participation in Environmental Research*. *World Agroforestry Journal* (Vol. Occasional). Nairobi.

Hyder, K., Townhill, B., Anderson, L. G., Delany, J., & Pinnegar, J. K. (2015). Can citizen science contribute to the evidence-base that underpins marine policy? *Marine Policy*, *59*, 112–120. https://doi.org/10.1016/j.marpol.2015.04.022

IEEE. (1991). *IEEE Standard for Developing Software Life Cycle Processes*. *IEEE Softwares* (Vol. 1997). New York.

IPCC. (2014). Climate Change 2014 Synthesis Report Summary Chapter for Policymakers. *Ipcc*, 31. https://doi.org/10.1017/CBO9781107415324

Irwin, A. (1995). *Citizen Science: A Study of People, Expertise and Sustainable Development*. (Routledge, Ed.), *Citizen Science: A Study of People, Expertise and Sustainable Development* (Vol. 10). New-York: Taylor and Francis e-Library.

Irwin, A., & Michael, M. (2003). *Science, social theory and public knowledge*. New York: Open University Press.

Jarrar, M., & Meersman, R. (2008). Ontology Engineering -The DOGMA Approach 1 Introduction and motivation. In *Advances in Web Semantics*. Berlin, Heidelberg: Springer-Verlag.

Johnson, M. F., Hannah, C., Acton, L., Popovici, R., Karanth, K. K., & Weinthal, E. (2014). Network environmentalism: Citizen scientists as agents for environmental advocacy. *Global Environmental Change*, *29*, 235–245. https://doi.org/10.1016/j.gloenvcha.2014.10.006

Jollymore, A., Haines, M. J., Satterfield, T., & Johnson, M. S. (2017). Citizen science for water quality monitoring: Data implications of citizen perspectives. *Journal of Environmental Management*, *200*, 456–467. https://doi.org/10.1016/j.jenvman.2017.05.083

Jones, D., Bench-Capon, T., & Visser, P. (1998). Methodologies for ontology development. *Conference of the 15th IFIP World …*, (April 2016), 20–35. https://doi.org/10.1.1.52.2437

Kaufman, A., Williams, R., Barzyk, T., & Hagler, G. (2016). Citizen science opportunities for monitoring air

quality. *Environmental, The U S Science, Citizen Monitoring, Quality Monitoring, Generation Air*.

Kim, S., Iglesias-Sucasas, M., & Viollier, V. (2013). The FAO Geopolitical Ontology: A Reference for Country-Based Information. *Journal of Agricultural & Food Information*, *14*(1), 50–65. https://doi.org/10.1080/10496505.2013.747193

Knights, K. (1976). Legal and Ethical Issues in data sharing. United Kingdom: United Kingdom Data Services. https://doi.org/10.1016/B978-0-323-01199-0.50225-5

Kolok, A. S., Schoenfuss, H. L., Propper, C. R., & Vail, T. L. (2011). Empowering Citizen Scientists: The Strength of Many in Monitoring Biologically Active Environmental Contaminants. *BioScience*, *61*(8), 626–630. https://doi.org/10.1525/bio.2011.61.8.9

Kotis, K., & Vouros, G. A. (2006). Human-centered ontology engineering: The HCOME methodology. *Knowledge and Information Systems*, *10*(1), 109–131. https://doi.org/10.1007/s10115-005-0227-4

Kraak, M. J., & Ormeling, F. (2011). *Cartography : visualization of spatial data*. Guilford Press. Retrieved from https://www.guilford.com/books/Cartography/Kraak-Ormeling/9781609181932/reviews

Ledermann, N., Schwartz, R., & Abd-El-Khalick, F. (2015). *Encyclopedia of Science Education*. *Encyclopedia of Science Education*. https://doi.org/10.1007/978-94-007-2150-0

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Legrand, M., & Chlous, F. (2016). Citizen science, participatory research, and naturalistic knowledge production: Opening spaces for epistemic plurality (an interdisciplinary comparative workshop in France at the Muséum national d'Histoire naturelle ["National museum of natural History"]). *Environmental Development*, *20*, 59–67. https://doi.org/10.1016/j.envdev.2016.10.002

Lemmens, R., Falquet, G., & Métral, C. (2016). Towards Linked Data and ontology development for the semantic enrichment of volunteered geo-information, 2–6.

Lieberman, J., Singh, R., & Goad, C. (2007). W3C Geospatial Ontologies. Retrieved September 5, 2017, from https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/

Lintott, C. (2017). Volunteers and professionals make real discoveries together. Retrieved October 24, 2017, from https://www.zooniverse.org/about

Liu, P., Hu, Y., Wang, X., & Liu, K. (2011). A methodology for domain ontology construction in information science. In *2011 International Conference on E-Business and E-Government (ICEE)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICEBEG.2011.5882759

López, F. (1999). Overview Of Methodologies For Building Ontologies. *Proceedings of the IJCAI99 Workshop on Ontologies and ProblemSolving Methods Lessons Learned and Future Trends CEUR Publications*, *1999*(2), 1–13. https://doi.org/10.1.1.39.6002

Lozano-Tello, A., & Gómez-Pérez, A. (2004). Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, *2*(15), 1–18. https://doi.org/10.1109/MC.2002.1046975

Luciano, J. S., Obrst, L., Stoutenburg, S., Cohen, K., & Standford, J. (2008). Ontology Evaluation : Methods and Metrics Ontology : A Key Technology for Knowledge Management, 1–19.

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. *Conservation Biology*, *30*(3), 447–449. https://doi.org/10.1111/cobi.12706

Mäkelä, J. M. (2006). The Impact of Spatial Data Quality on Company's Decision Making. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *34*.

Margules, C. R., Pressey, R. L., & Williams, P. H. (2002). Representing biodiversity: Data and procedures for identifying priority areas for conservation. *Journal of Biosciences*, *27*(4), 309–326. https://doi.org/10.1007/BF02704962

Masolo, C., & Borgo, S. (2005). Qualities in formal ontology. *Foundational Aspects of Ontologies (* …, 15. Retrieved from https://www.uni-koblenz-landau.de/koblenz/fb4/forschung/publications/fachberichte/fb2005/rr-9-2005.pdf#page=8

Matthews, S. A., & Parker, D. M. (2013). Progress in spatial demography. *Demographic Research*, *28*(February), 271–312. https://doi.org/10.4054/DemRes.2013.28.10

McGarigal, K. (2001). Concepts of Scale. *Landscape Ecology*. North Carolina: Duke University.

Meentemeyer, V. (1989). Geographical perspectives of space, time, and scale. *Landscape Ecology*, *3*(3–4), 163–173. https://doi.org/10.1007/BF00131535

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285–290. https://doi.org/10.1890/110278

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An online lexical database. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

Mustaffa, S., Ishak, R. Z., & Lukose, D. (2012). Ontology model for herbal medicine knowledge repository. *Communications in Computer and Information Science*, *295 CCIS*, 293–302. https://doi.org/10.1007/978-3-642-32826-8_30

NCBO. (2012). GeoSpecies Ontology - Summary | NCBO BioPortal. Retrieved December 19, 2017, from https://bioportal.bioontology.org/ontologies/GEOSPECIES

NCBO. (2017). Vertebrate Taxonomy Ontology - Summary | NCBO BioPortal. Retrieved December 19, 2017, from https://bioportal.bioontology.org/ontologies/VTO

NeoGeo. (2012). NeoGeo Spatial Ontology. Retrieved December 19, 2017, from http://geovocab.org/spatial

Newell, A. (1981). *The knowledge level. Artificial Intelligence* (Vol. 18). https://doi.org/10.1.1.103.3321

Nieland, S., Moran, N., Kleinschmit, B., & Förster, M. (2015). An ontological system for interoperable spatial generalisation in biodiversity monitoring. *Computers and Geosciences*, *84*, 86–95. https://doi.org/10.1016/j.cageo.2015.08.006

Nieto, M. (2003). An overview of ontologies. *Universidad De Las Américas Puebla, Interactive and Cooperative Technologies Lab*, (March). https://doi.org/10.1016/j.acalib.2010.11.015

Nov, O., Arazy, O., & Anderson, D. (2011). Technology-Mediated Citizen Science Participation: A Motivational Model. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, (July), 249–256. https://doi.org/10.1145/1940761.1940771

Oberle, D. (2014). How ontologies benefit enterprise applications. *Semantic Web*, *5*(6), 473–491. https://doi.org/10.3233/SW-130114

OGC. (1999). *Simple features specification for SQL Revision 1.1. Open GIS*. Virginia.

Oliveira, P., Rodrigues, F., & Henriques, P. (2006). An ontology-based approach for data cleaning. *Proceedings of the 2006 International Conference on Information Quality, ICIQ 2006*, *6*, 0–10. https://doi.org/10.1.1.67.7586

OS. (2015). Ordnance Survey Ontologies. Retrieved December 19, 2017, from http://data.ordnancesurvey.co.uk/ontology

Oussous, A., Benjelloun, F.-Z. Z., Ait Lahcen, A., & Belfkih, S. (2017). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2017.06.001

Ovchinnikova, E. (2012). Natural Language Understanding and World Knowledge. *Integration of World Knowledge for Natural Language Understanding*, (1996), 15–37.

Pan, X., & Pohl, J. G. (2009). Conveyance Estimator Ontology: Conceptual Models and Object Models. *Proceedings of InterSymp-2009: Baden-Baden, Germany*, (Gruber 2008), 1–12.

Pavlic, T. P., & Pratt, S. C. (2013). Participating in Online Citizen Science: Motivations as the Basis for User Types and Trajectories. *Human Computation*, 911–960. https://doi.org/10.1007/978-1-4614-8806-4_74

Perego, A., & Lutz, M. (2015). ISA Programme Location Core Vocabulary. Retrieved December 19, 2017, from https://www.w3.org/ns/locn

Pérez, A., Baonza, M. D. F., & Villazón, B. (2008). Neon methodology for building ontology networks: Ontology specification. *Methodology*, (February), 1–18. https://doi.org/10.1016/j.landurbplan.2011.04.007

Perez, A. G., & Benjamins, V. R. (1999). Overview of Knowledge Sharing and Reuse Components : Ontologies and Problem-Solving Methods. *IJCAI-99 Workshop on Ontologies and Problem-Solving Method (KRR5)*, 1–15.

Pettibone, L., Vohland, K., Bonn, A., Richter, A., Bauhus, W., Behrisch, B., … Ziegler, D. (2016). Citizen science for all, 56.

Pocock, M. J. O., Chapman, D. S., Sheppard, L. J., & Roy, H. E. (2014). 012-x-013-Choosing and using

citizen science: A guide to when and how to use citizen science to monitor biodiversity and the environment. *Centre for Ecology & Hydrology*. Retrieved from https://www.ceh.ac.uk/sites/default/files/sepa_choosingandusingcitizenscience_interactive_4web_final_amended-blue1.pdf

Poli, R. (2003). Descriptive, Formal and Formalized Ontologies. *Husserl's Logical Investigations Reconsidered*, *48*(1), 183–210.

Rajpathak, D., & Chougule, R. (2011). A generic ontology development framework for data integration and decision support in a distributed environment. *International Journal of Computer Integrated Manufacturing*, *24*(2), 154–170. https://doi.org/10.1080/0951192X.2010.531291

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, *36*, 1–22. https://doi.org/10.1016/j.websem.2016.01.001

Robertson, C. (2015). Whitepaper on Citizen Science for Environmental Research. *Spatial Research*. Ontario: Wilfrid Laurier University. https://doi.org/10.13140/RG.2.2.27252.14724

Roman, L. A., Scharenbroch, B. C., Östberg, J. P. A., Mueller, L. S., Henning, J. G., Koeser, A. K., … Jordan, R. C. (2017). Data quality in citizen science urban tree inventories. *Urban Forestry and Urban Greening*, *22*, 124–135. https://doi.org/10.1016/j.ufug.2017.02.001

Sajja, P. S., & Akerkar, R. (2010). Knowledge-Based Systems for Development. *Advanced Knowledge Based Systems: Model, Applications & Research*, *1*, 1–11. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Success+Factors+in+Implementing+Knowledge+Based+Systems#0

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co Design*, *4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Schade, S., & Tsinaraki, C. (2016). *Survey Report: Data Management in Citizen Science Projects*. *JRC Technical Reports*. https://doi.org/10.2788/539115

Schnoor, J. L. (2007). *Citizen science. Environmental science & technology* (Vol. 41). https://doi.org/10.1353/eco.2007.0002

Schooley, J. (2017). *Introduction to Botany*. *Mycorrhiza*. North Dakota: Minot State University.

Schulz, S., Stenzhorn, H., & Boeker, M. (2008). The ontology of biological taxa. *Bioinformatics (Oxford, England)*, *24*(13), i313-21. https://doi.org/10.1093/bioinformatics/btn158

Schwartz, J. T., Cantone, D., & Omodeo, E. G. (2011). *Computational Logic and Set Theory: Applying formalized logic to analysis*. *Computational Logic and Set Theory* (1st ed.). London: Springer-Verlag. https://doi.org/10.1007/978-0-85729-808-9

SciStarter. (2017). Drug discovery from your soil on SciStarter. Retrieved October 19, 2017, from https://scistarter.com/project/1164-Drug-discovery-from-your-soil#sthash.9SV7HW3X.dpbs

Sintek, M., Buitelaar, P., & Olejnik, D. (2007). A Formalization of Ontology Learning From Text. *Knowledge Management*. Kaiserslautern: DFKI GmbH Language Technology Department.

Socientize, P. (2013). Green paper on Citizen Science. Citizen Science for Europe: Towards a society of empowered citizens and enhanced research (pp. 1–54). https://doi.org/http://dx.doi.org/10.1126/science.1156895

Staab, S., & Stuckenschmidt, H. (2006). *Semantic web and peer-to-peer decentralized management and exchange of knowledge and information*. *Semantic Web and Peer-to-Peer: Decentralized Management and Exchange of Knowledge and Information*. https://doi.org/10.1007/3-540-28347-1

Staab, S., & Studer, R. (2007). Handbook on Ontologies. *Decision Support Systems*, 654. https://doi.org/10.1007/978-3-540-92673-3

Staroch, P. (2013). *A Weather Ontology for Predictive Control in Smart Homes*. Vienna University of Technology.

Stuckenschmidt, H., & Visser, U. (1999). Ontologies for geographic information integration. *Proceedings of Workshop …*, 1–17. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.6794&rep=rep1&type=pdf

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Thessen, A. E., & Patterson, D. J. (2011). Data issues in the life sciences. *Me-Infrastructures for Data Publishing*

in Biodiversity Science, *150*, 15–51. https://doi.org/10.3897/zookeys.150.1766

Tinati, R., Luczak-Roesch, M., Simperl, E., & Hall, W. (2017). An investigation of player motivations in Eyewire, a gamified citizen science project. *Computers in Human Behavior*, *73*, 527–540. https://doi.org/10.1016/j.chb.2016.12.074

Tom, H., & Roswell, C. (2009). Standards Guide ISO/TC 211 GEOGRAPHIC INFORMATION/GEOMATICS. *ISO/TC 211 Advisory Group on Outreach*, (June), 98. Retrieved from http://www.isotc211.org/Outreach/ISO_TC_211_Standards_Guide.pdf

Turner, A. J. (2006). Introduction to Neogeography. *OReilly*. O'Reilly Media.

UNFCCC. (2007). Climate Change: Impacts, Vulnerabilities and Adaptation in Developing Countries. *United Nations Framework Convention on Climate Change*, 68. https://doi.org/10.1029/2005JD006289

Uschold, M. (1996a). Building Ontologies : Towards a Unifed Methodology, (September).

Uschold, M. (1996b). Building Ontologies : Towards a Unified Methodology. In University of Edinburg (Ed.), *16th Annual Conference of the British Computer Societ y Specialist Group on Expert Systems*. United Kingdom.

Uschold, M., Uschold, M., Healy, M., Williamson, K., Clark, P., Healy, M., … Woods, S. (1998). Ontology reuse and application. *Formal Ontology in Information Systems*, *179*(January 2000), 192.

USGS. (2010). Did You Feel It? Retrieved October 19, 2017, from https://earthquake.usgs.gov/data/dyfi/

USGS. (2017). Seismic Hazard Maps and Site-Specific Data. Retrieved November 1, 2017, from https://earthquake.usgs.gov/hazards/hazmaps/

Varzi, A. C. (2011). On doing ontology without metaphysics. *Philosophical Perspectives*, *25*, 407–423. https://doi.org/10.1111/j.1520-8583.2011.00222.x

Visser, U., Visser, U., Stuckenschmidt, H., Schuster, G., & Vögele, T. (2002). Ontologies for geographic information processing. *Computer and Geosciences*, *28*(1), 103; 103-117; 117.

Viswanathan, G., & Schneider, M. (2011). On the Requirements for User-Centric Spatial Data Warehousing and SOLAP. *Processing*, *6637*, 144–155. https://doi.org/10.1007/978-3-642-20244-5_14

Viswanathan, G., & Schneider, M. (2013). User-centric spatial data warehousing: a survey of requirements and approaches. *Int. J. Data Mining, Modelling and Management*, *x*(x), 22. https://doi.org/10.1504/IJDMMM.2014.066764

W3C. (2007). GeoRSS in RDF. Retrieved October 6, 2017, from http://www.georss.org/rdf_rss1.html

W3C. (2014). *RDF Schema 1.1. World Wide Web Consortium*. Retrieved from http://www.w3.org/TR/2014/REC-rdf-schema-20140225/

Willig, M. R., & Presley, S. J. (2017). Biodiversity and Disturbance. In *Encyclopedia of the Anthropocene* (pp. 45–51). Elsevier. https://doi.org/10.1016/B978-0-12-809665-9.09813-X

Zorica, N.-B., Budhathoki, N., & Bruce, B. C. (2010). An Interdisciplinary Frame for Understanding Volunteered. *Geomatica*, *64*(1), 2010–2010.

APPENDIX

Figure 0-1: Preview of The Raw Combined Dataset in RDF. : Source: Author

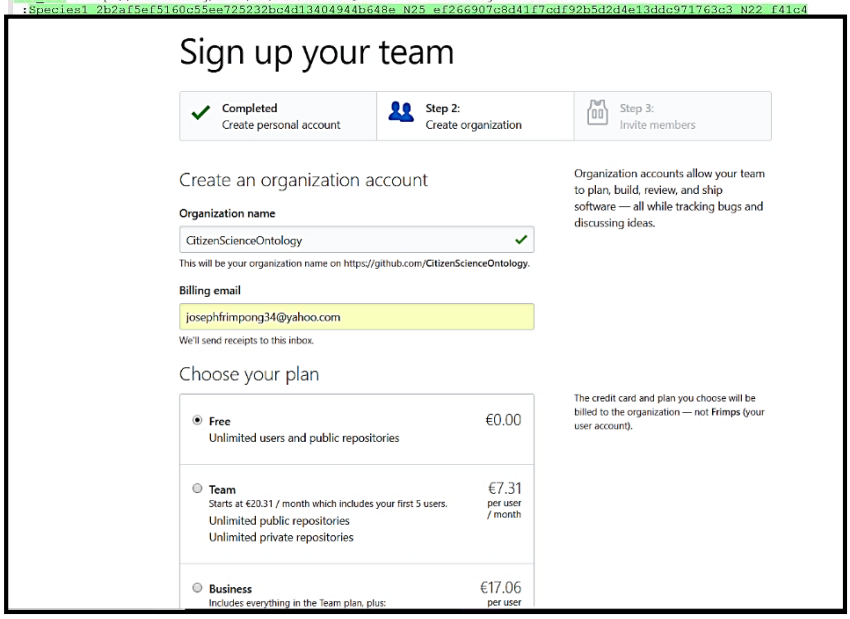

Figure 0-2: Account Sign UP GitHub Platform: Source: Author

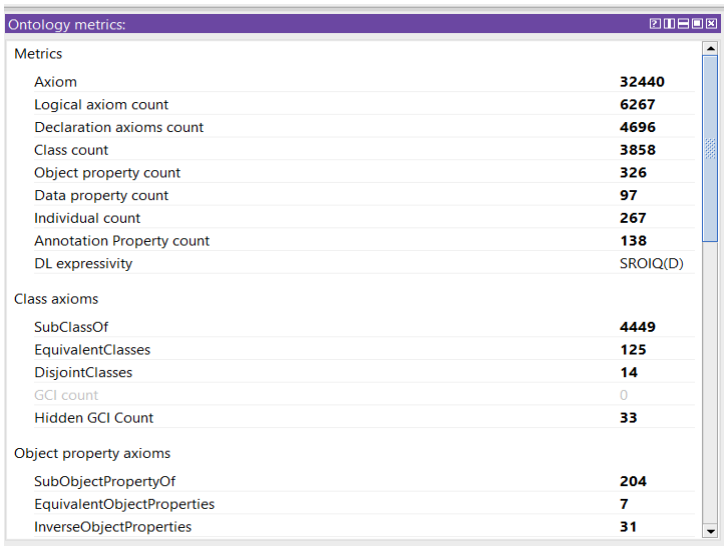

Figure 0-3: Ontology Metrics for Metric Sute Testing: Source: Author

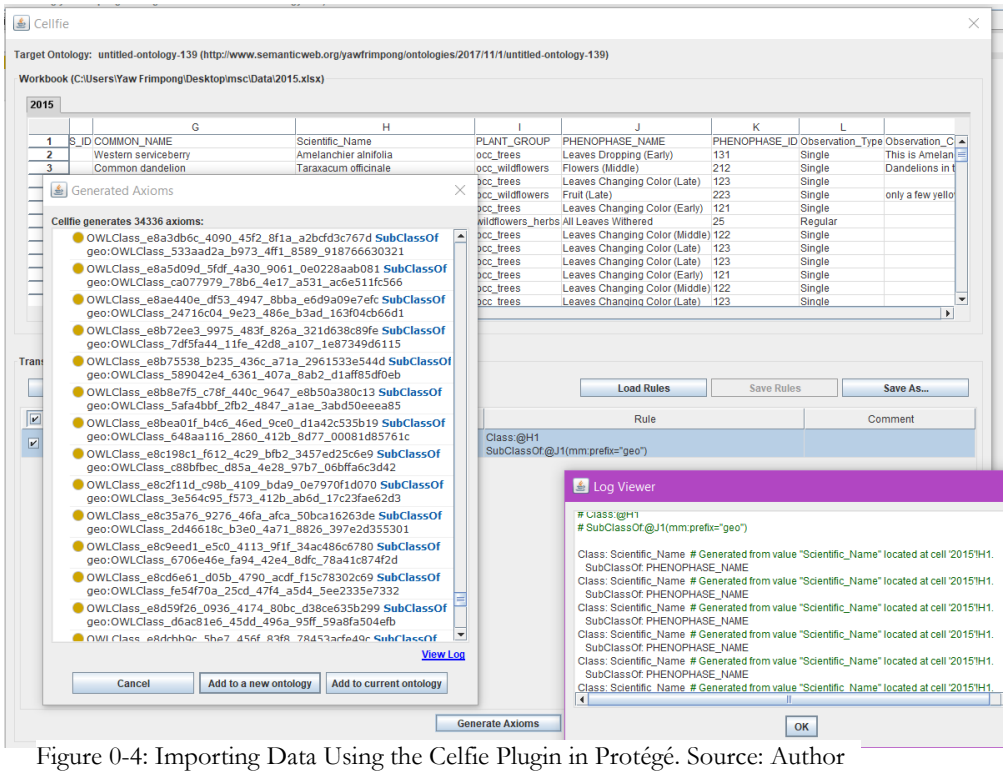

Figure 0-4: Importing Data Using the Celfie Plugin in Protégé. Source: Author

```
#Importing Modules
import rdflib
from rdflib.graph import Graph, URIRef
import matplotlib.pyplot as plt
import gzip
#Loading and Serilizin Combined Dataset
from gastrodon import LocalEndpoint,one,QName
#g = rdflib.ConjunctiveGraph()
g = Graph()
#g.parse("C:/publish/RDF/OWL.ttl",format="ttl")
g.parse("C:/publish/RDF/WSP1WS8.ttl",format="ttl")
len(g)
e=LocalEndpoint(g)
#qUERY formulation
properties1=e.select("""
SELECT ?o ?o1 ?o2
WHERE {
?s <http://www.semanticweb.org/
\yawfrimpong/ontologies/untitled-ontology-13#FoundOn>  ?o .
?s <http://www.semanticweb.org/yawfrimpong/ontologies\
/untitled-ontology-13#HasLat> ?o1 .
?s <http://www.semanticweb.org/yawfrimpong/ontologies\
/untitled-ontology-13#HasLong> ?o2 .
?s <http://www.semanticweb.org/yawfrimpong/ontologies\
/untitled-ontology-13#HasLong> "San Francisco" .
?s <http://www.semanticweb.org/yawfrimpong/ontologies\
/untitled-ontology-13FeedOn#> "Forest" .
}
    """)
properties1
#Saving and Printing Result
print(properties1)
```

Figure 0-5: Code for selecting the different land cover classes. Source: Author

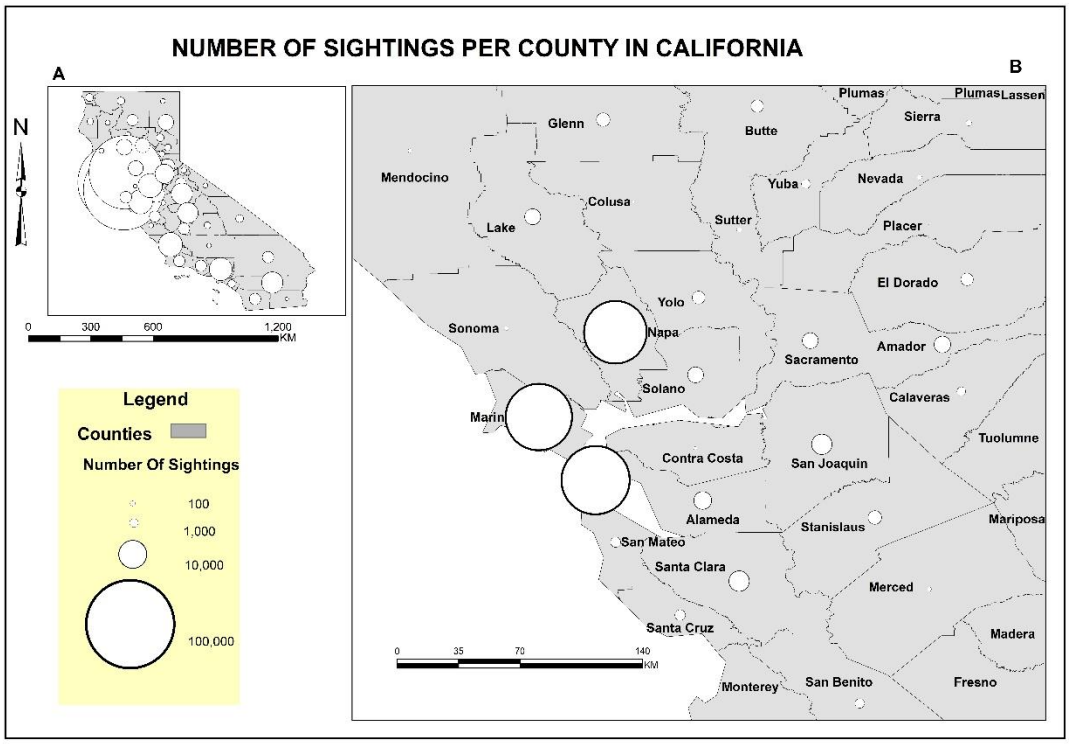

Figure 0-6: Results of The Query (Q1.1) Visualised in ArcMap. **A** show the list of Counties with their proportion. **B** shows a zoomed in version of the **A.** Source: Author

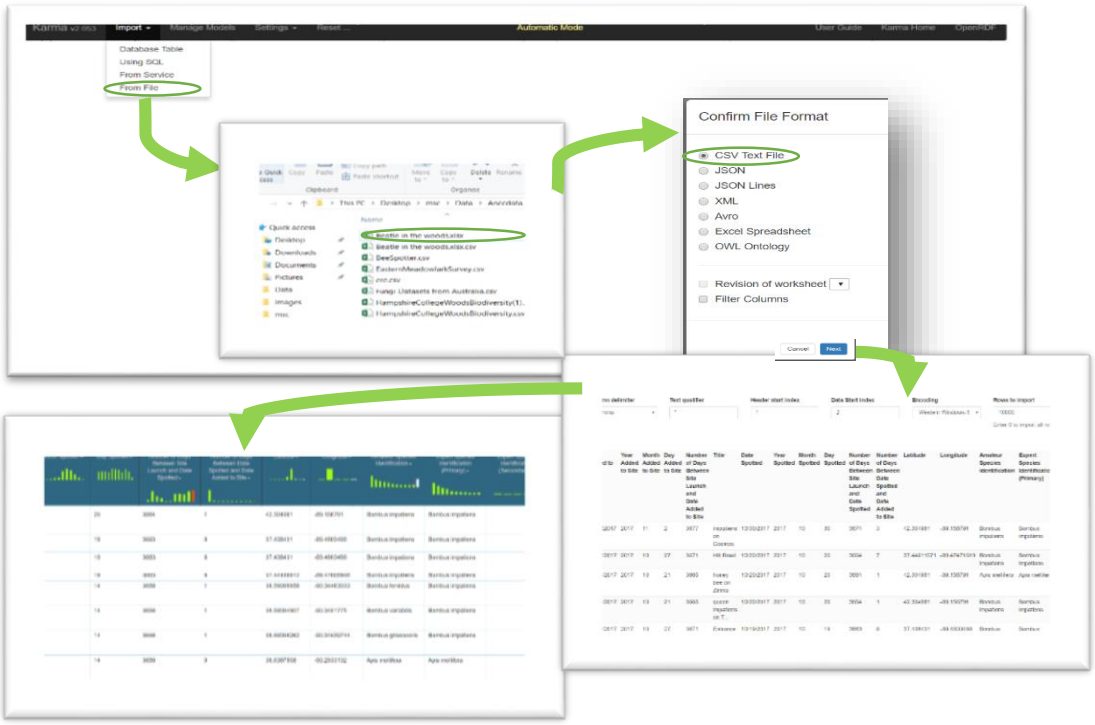

Figure 0-7: Importing Different Datasets into the Karma Environment. Source: Author

```
import rdflib
from rdflib.graph import Graph, URIRef
import matplotlib.pyplot as plt
import gzip
import numpy as np
import pandas as pd
from gastrodon import LocalEndpoint,one,QName

#g = rdflib.ConjunctiveGraph()
g = Graph()

g.parse("C:/publish/RDF/OWL.ttl",format="ttl")
len(g)

e=LocalEndpoint(g)
properties1=e.select("""

SELECT ?o ?o2
 {
        ?s <http://www.semanticweb.org/yawfrimpong/ontologies/\
        untitled-ontology-13#FeedOn> "GrassLand" .
        ?s <http://www.semanticweb.org/yawfrimpong/ontologies/\
        untitled-ontology-13#HasLat> ?o .
        ?s <http://www.semanticweb.org/yawfrimpong/ontologies/\
        untitled-ontology-13#FoundIn> "San Francisco" .
        ?s <http://www.semanticweb.org/yawfrimpong/ontologies/\
        untitled-ontology-13#HasName> "Owl" .
          }  """)
properties1
file = open("GrassLand.txt","w")
data = properties1.values
for row in data:
    file.write(str(row[0])+" "+str(row[1])+"\n")
file.close()
```

Figure 0-8: Code Interface for selecting Diffrent Classes for SPARQL Reasoning. Source: Author

Table 0-1: : Result from Semantic Score (Metric Suite Testing): Source: Author

6	Conservation	Active management of the biosphere to ensure the survival of the maximum diversity of species and the maintenance of genetic variability within species. It includes the maintenance of biosphere function, e.g. nutrient cycling and ecosystem function.	An occurrence of improvement by preventing loss or injury or other change	3
7	habitat	A place or type of site where an organism or population naturally occurs	The type of environment in which an organism or group normally lives or occurs	1
8	Species	An interbreeding group of organisms that are reproductively isolated from all other organisms, although there are many partial exceptions to this rule in particular taxa	Taxonomic group whose members can interbreed	2
9	City	An international agreement between governments which aims to ensure that international trade in specimens of wild	A large and densely populated urban area; may include several independent administrative districts	3

		animals and plants does not threaten their survival		
10	Population	A group of individuals of the same species, occupying a defined area and usually isolated to some degree from other similar groups	The people who inhabit a territory or state	5
11	Taxon	A taxon (plural: taxa), or taxonomic unit, is a unit of any rank (i.e. kingdom, phylum, class, order, family, genus, species) designating an organism or a group of organisms	Animal or plant group having natural relations	1
12	Nomenclature	A systematic naming of things or a system of names or terms for things. In classification, nomenclature involves a systemic naming of categories or items	A system of words used to name things in a particular discipline	1
13	Landscape	An area of land that contains a mosaic of ecosystems, including human-dominated ecosystems.	An expanse of scenery that can be seen in a single view	4
14	Extinction	The condition that arises from the death of the last surviving individual of a species, group or gene globally or locally.	No longer active; extinguished	6
15	Diversity	The variety and relative abundance of different entities in a sample	Noticeable heterogeneity	2
16	Ecology	A branch of biology which addresses the relationships between living organisms and their environment. Ecology can be addressed at some scales; it also includes the relationships of a particular organism with its environment	The environment as it relates to living organisms	2
17	Environment	The totality of all the external conditions affecting the life, development and survival of an organism	The totality of surrounding conditions	2
18	Watershed	The land area that drains into a particular watercourse or body of water. Sometimes used to describe the dividing line of high ground between two catchment basins	A ridge of land that separates two adjacent river systems	3
19	Seagrass	A group of flowering plants found in marine or estuarine waters that tend to develop extensive underwater meadows.	Any of various seaweeds that grow underwater in shallow beds	1
20	Assemblage	A collection of species inhabiting a given area, the interactions between the species, if any, being unspecified.	A group of persons together in one place	4

Table 0-2A: Some Details on Existing Effort Towards Spatial Ontologies Design. Source: Author

Ontology	General information	Spatial encoding strategies	Review
Ordnance survey ontologies .	The ontology is designed by the Ordnance Survey of Great Britain. Ordinance survey provides up-to-date and accurate geographic information to stakeholders in Great Britain The ontologies consist of five different ontologies describing five different spatial components in the geospatial information domain. These five ontologies include the postcode ontology, the administrative geography and local voting area ontology, the spatial relation ontology, the geometry ontology and the 50k gazetteer ontology. These set of ontologies were designed to capture and express spatial information such as postcodes, abstract spatial geometries, administrative and voting area of Great Britain. The ontology has spatial concepts such as country, which describes the geographic position of a sovereign state. The five different ontologies are expressed to capture a different aspect of geographic information. The geometry ontology expresses the abstract geometries of spatial objects. The spatial relation ontology also describes spatial relations such as contain, disjoints, equals which serves as a spatial predicate to determine the relationships between two corresponding geometries.	The OS ontology models feature as sub-ontology in each subdomain. Examples of features include (transport, administrative unit, hydrography, etc.). Geometries are modelled as literals using the Geographic Markup Language (GML) datatypes in RDF.	The Ordinance Survey ontology stands out to be among the well-developed and well-structure ontologies on spatial relations for business entities and stakeholders to use. It contains about 109412167 triples with Talis API as the data asses. Datasets used are obtained directly from the Ordinance Survey Datasets. *Source: (OS, 2015)*
NeoGeo Ontology.	The NeoGeo ontology is a geospatial ontology which models all geospatial information into feature and geometry. From the NEOGEOVOCAMP group, they set a distinction between a geometry and features is a geometry is defined as anything that has a geometric shape. Whiles, a feature is anything with spatial extent. The NeoGeo ontology provides spatial:Feature and geom:Geometry the two main classes. The relation between a spatial:Feature class and a geom:Geometry class is the property geom:geometry.	The NeoGeo vocabulary has two basic classes, these two classes are the Spatial:Feature and the Geom:Geometry. The spatial:Features models geographic features as anything capable of holding spatial relation. It also models Spatial Geometries as bounding boxes composed of four-line segment defining a rectangle	The ontology is widely re-use and considered as one of the most effective representations of geographic information on the semantic web. It has an exponential growth of triples due to

		the geometry is enclosed in. (geom: Geometry)	multiple re-use capabilities. *Source: (NeoGeo, 2012)*
Geopolitical ontology	The Food and Agriculture Organisation (FAO) developed the geopolitical ontology. This ontology is developed to support information exchange and sharing in a regulated system among frameworks and organisations superintending data about countries and its territories. It has been expressed in different natural languages. Examples of the languages include English, French, Spanish, Arabic, Chinese, Russian and Italian. The ontology is populated with data from the United Nation (UN), FAO and other internationally recognised data sources. It also tracks changes to historical geopolitical datasets historically to determine the trustworthiness of the data holders and data providers.	The FAO ontology modules feature as spatial entities that can be described by spatial relations. In their interest, most features are countries. The ontology also modules spatial geometries as bounding boxes. These bounding boxes are defined using their minimum Longitude, minimum Latitude, maximum Longitude, maximum Latitude of the geometry.	The FAO geopolitical ontology has been codded in multiple languages with the aim of enhancing communication among all associated partners. The ontology has links to different datasets on different platforms such as the DBpedia and the current LOD cloud. Source: (Kim et al., 2013)
ENERGIC (VGI) Ontology	The VGI ontology holds concepts used in Volunteered Geo-Information which can be considered as an alias of Citizen Science for the geographic information domain. These concepts are derived from keywords which have been extracted from the VGI Handbook. The concepts were edited on a web protégé platform and can be assessed through RDF in the Web Protégé environment.	The domain of VGI is grouped into different top categories which include Data, Tool, Project, Method, Person, Publ ication, Organization, Event.	There are no records of reuse in this ontology. However, the platform provides links to other useful resources. **Source**: (Lemmens et al., 2016)
ISA Programme Location Core Vocabulary	The ISA Programme Location Core Vocabulary is designed to facilitate the publication of data that is interoperable with EU INSPIRE Directive. The vocabulary captures a different set of classes and adequate properties for describing these classes. The classes and their relationships provide an adequate description of places by its name, address and geometry. The ontology is added to the W3C space as a registered organisation vocabulary.	The ISA defines geometry based on the initial encoding contained in the entities. Examples include Geometries as class, literal or geocoded URI. The class geometry is expressed to identify locations as point, lines and polygons.	No Records on triples reported. However, the vocabulary serves as a framework for defining spatial geometries and relations. **Source**: (Perego et al., 2015)

W3C Geospatial ontology	The geospatial incubator group designs the W3C Geo vocabulary. It aims at providing namespaces for defining both longitude and latitudes and other spatial entities using the WSG84 as a reference datum. It addresses most questions relating to geoinformation resources and their properties.	The W3C geospatial ontology uses the georss encoding strategy for defining spatial relation and spatial classes in the W3C geo vocabulary. This vocabulary uses LOD cache as its SPARQL endpoint. It models geometry as latitude and longitude of spatially objects.	The W3C Geo vocabulary uses around 21 different datasets with approximately 15543 105 number of triples. it is among the most widely used geospatial ontology. **Source**: (Lieberman et al., 2007)

Table 0-3: Means for merging the selected classes (Connecting Strategy). Source Author

Questions	Classes	Relation	Mappings Examples	Examples of Individuals	Datasets	Connecting Strategy	Remarks
SET 1							
Q1.1.1 Which region is the highest reported number of species?	County and Species	*FoundIn*	Owl-1 *FowndIn* (Marine_County), Waterfowl-1 *FoundIn* San Francisco County	Owl-1, Owl-2, Drag-1 moreover, all possible species in the datasets with the same county	All datasets from the sightings in the use case	Triples are formed with individuals as subject, and CountyName as (object:value) joined with the relation *FoundIn*	The SPARQL query selected counted all species based on their county and selected most occurring species observation the most occurring county. The Result could have been done with a traditional GIS. However, the question serves as an input to question two. It also emphases the capabilities of the ontology been able to perform other general operations

Q1.1.2 What are the land use and land cover characteristics of those regions?	Species Dragonfly, County, LandCover, LandUse.	*FoundOn*	Owl-1 *FoundOn* Forest Area Owl-2 *Eat* Drag-1 *FoundOn* SwampyArea	Owl-1, Owl-2, Drag-1, Drag-2	All the datasets	Forming triples such as Drag-1 *FoundOn* SwampyArea and SwampyArea *SubClassOf* Forest. Forest *HabitatFor* Dragonfly This mapping considers the N-ary mapping strategy where different individuals are mapped to each other depending on the defined relations in the ontology.	The SPARQL query selected the different land use and land cover information according to the selected county from Query 1. Therefore, query 1 served as the input for the selection of the different land use and land cover information. This species, their Latitude and longitude values of the choosing county for the different species. These selections are to provide spatial information of the areas reported using the latitude and longitude. Moreover, the proposed land classes that need validation can be plugged into the query to find all possible location of the selected land class according to the dataset.
Q1.3 Are there any risks of natural	Risk, Species and County	*HasRisk*,	If Owl-1 is a selected member of the county from the previous query to determine the risk of the	List of all species in the selected county from the section 1 question 1.	All the dataset for this use case	The relation *HasRisk* in the ontology is mapped to the class risk. This mapping is to check the number of possible risk in the Areas. The data property *HasLat* and *HasLong* is referenced to the geolocation mapping	A SPARQL query is to select all locations of species in the selected county. Followed by a traditional GIS analysis of the relationships to selected areas.

disaster reported in these areas?			selected county.			in the W3C Geo relation (Geo:lat and Geo:Long) respectively. Therefore, both proximity relation such as within can be used to confirm the distance of the areas with Risk-Prone Areas. (More detail will be proposed for future work .)	
Q1.4 What are the different land classes that can be obtained from the sightings?	Species Dragonflies, County, Forest	*FoundOn* *Eat*	Owl-1 *FoundOn* SwampyArea Owl-2 *Eat* Drag-1 *FoundOn* SwampyArea	Owl-1, Owl-2, Drag-1, Drag-2	All the datasets	Forming triples such as Drag-1 *FoundOn* SwapyArea and SwapyArea *SubClassOf* Forest. Forest *HabitatFor* Dragonfly This mapping considers the N-ary mapping strategy.	The SPARQL query is to select species, Lat values and Long value of the chosen county with the different species form Query Q1.1. This selection is to provide spatial information of the areas reported using the Lat and Long information. Individuals grouped in the same area will have same environmental characteristics. Therefore, the well-sampled area was determined by the most reported species and the areas they are reported.

Q1.5	Owl, LandCover	*FoundOn* *FoundIn* and *HasLocation*	Owl-1 *HasLocation* (County) Owl-2 *FoundIn* Rainforest	All Instance of Owl	Owl Dataset and landcover information	Forming triples with the following relations *FoundIn*, *FoundOn*, *HasLocation*	The SPARQL query selected all the values from the Owl sightings using the relations *HasLocation*, *FoundOn* and *FoundIn* The selection will collect Lat and longitude values of Owl species with the value from the selected relations. This selected information will then serve as the validator for the user input landcover classes
SET 2							
Q2.1	Waterfowl and County	*HasLocation, HasLat, HasLong* *HasDate*	Waterfowl-1 *FoundIn* County Waterfowl-2 *HasLat* (Value) Waterfowl-3 *RecordedAt* (Date)	All instance of waterfowls	Datasets on Waterfowls	Individuals will be connected using their corresponding data properties. *FoundIn*, *HasLat*, *HasLong* and *HasLat*	The SPARQL query select is to select all reported waterfowls and ordered them by Date.
Q2.2	Waterfowl	*HasLong,*	Waterfowl-1 *FoundOn* (Landuse type: Value)	All instance of waterfowls	Datasets on	Individual will be connected with the characteristics using the value pairs from the	A SPARQL query is to select all the characteristics of the Land use using relation involving these areas.

	County and LandCover	*HasLat, HasLocation, FoundIn, FoundOn*			Waterfowls, County information, Landcover types	corresponding data properties (Mapping relation)	The value pair with the relations *FoundIn* and *FoundOn*.
Q2.3.	Species	*FoundOn FoundIn*	Waterfowl-1 *FoundOn* (Landuse type:Value) Waterfowl-2 *FoundOn* (Landuse type:Value)	Instances of all species in the selected dataset.	Dataset on all species	Individuals will be connected using their corresponding data and object properties. *FoundOn, HasLat, HasLong and FoundIn*	A SPARQL query is to select and process all species within the selected location specific distance from the selected species in section 2 query 2.
SET 3							
Q3.1[31]	Birds, LandUse LandCover,	*Breeds, ConsumedBy, Consumes*	Sunbird *Is-A-type-Of* Bird, Sunbird-1 *Consumes* Nectar, Swamps *Breeds*	Instances of the different types of birds and insects in the dataset	Datasets on the different land cover types, Birds, and other insects	The connection with different individuals is based on the data and object properties on the datasets.	A GeoSPARQL query is to select areas with Most species of birds and insect. However, doesn't fall under the restricted areas.

[31] The different classes here used in connecting the different classes during the data integration stage. Therefore, they were not used in the used directly in the query.

	Nectar, Plants and Insects	*Is-A-Type-Of* and *Contains*	DragonflyNymph				Note[32]

[32] The query is not fully implemented due to the use of GeoSPARQL query.

Table 0-4: Translating Competency questions to Queries using Set Notations and Logics. Source: Author

Question	Explanation	Reasoning	Intermediate	Mathematics	Query
SET 1					
Q1.1 Which region is the highest reported number of species?	Which region of the combined datasets has more sightings? Retrieve names of all counties and the reported number of species per each county.	For all Subject (Species) select the objects and subject with the relation FoundIn. Group the result by County name and count per County.	For all species in the datasets ($\forall S \in N$) select ($\mid$) [objects (O_i) mapped with the relation FoundIn ($\leftarrow (FoundIn)_i \rightarrow$) to such species($S_i$)] Save results as A ($\Rrightarrow A$). Group A according to Distinct objects ($\curvearrowright ((O_a \rightarrow O_b)$ and select Object group (O_b) with the maximum number ($\mid max(O_b)$) and save final results as X ($\Rrightarrow X$).	$\{\forall S \in N \mid [[O_i (\leftarrow (FoundIn) \rightarrow S_i] \Rrightarrow A] \curvearrowright O_a \rightarrow O_b \mid max(O_b)\} \Rrightarrow X$	SELECT ?o (COUNT(?o) as ?oCount) WHERE { ?s cs:FoundIn ?o .} GROUP BY ?p ORDER BY DESC(?oCount) Limit 1
Q1.2 What are the land use and land cover characteristics of those regions?	Retrieve the reported land use types for the for the result of (Q1.1)	For all subject mapped to object in X, select the objects these subject maps to using the relation FoundOn. Where object X is the county with the highest number of species	For all species in Query 1 ($\forall S \in X$) select ($\mid$) [objects (O_i) mapped with the relation FoundOn ($\leftarrow (FoundOn)_i \rightarrow$) to such species($S_i$)]	$(\forall S \in X) \mid [O_i (\leftarrow (FoundOn) \rightarrow) \rightarrow S_i]$	SELECT DISTINCT ?o { ?s ?p "San Francisco" . ?s cs:FoundOn ?o . }

Q1.3 Are there any risks of natural disaster reported in these areas?	Retrieve the type of risk associated with the Selected county in Q1.1.	Select the object mapped to the subject X with the relation (Predicate) HasRisk.	For County in Query 1 (C ∈ X) select (|) [objects (O_i) mapped with the relation HasRisk (← (HasRisk)$_i$ →) to such County(C_i)]	(C ∈ X) | O_i [O_i (← (HasRisk)$_i$ →) → C_i]	SELECT DISTINCT ?o { ?s ?p " San Francisco" . ?s cs:HasRisk ?o . }
[33]Q1.4. Where are the locations of Owl sightings?	Retrieve all values that are mapped to subject Owls using the relations (Predicate) HasLat and HasLong. Print the result into a table with columns longitude and latitude	For all subject describing Owls, select the HasLong and HasLat Values	For all species in the dataset (∀ S ∈ N) select (|) [Subject (S_i), Object ($O_{i,j}$) mapped with the relations HasLat (←(HasLat)$_i$ →) and HasLong (← (HasLong)$_j$ →) to such species(S_i)]. Save results as B (⇛ B). Select from result B Species (Si) with relation HasName and the value Waterfowls(W) (B | S_i (←(HasName)$_k$ →W).	{(∀ S ∈ N) | $O_{i,j}$ [$O_{i,j}$ (←(HasLat)$_i$ →) ∧ (← (HasLong)$_j$ →) → S_i]}⇛ B | S_k [S_k(←(HasName) →)→W)]}	SELECT ?s ?o1 ?o2 WHERE { ? s cs:HasName "Owl" . "Owl" cs:HasLat ?o1 . "Owl" cs:HasLong ?o2 . }
Q1.5[34] Are the sightings reported Forest cover as	Using the assumption that Owls are found in forest habitats. This query checks if for information from user inputs if the	For all Subject (Owl species), select distinct the values mapped with the relation FoundOn.	For all species in the dataset belonging to the Owl class (∀ S ∈ OW) select (|) [Distinct Objects (O_i) mapped with the relation FoundOn (←	(∀ S ∈ OW) | S_i,O_{jk} [O_{ijk} (← (*FoundOn*)$_i$ →) ∧ (← (HasLong)$_j$ →) ∧ (←	SELECT DISTINCT ?s ?o1 ?o2 WHERE { ?s cs:HasName "Owl" . ?s ?o "Forest" .

[33] Not Completed

[34] The query results obtained will serve as a reference for validating the different land information input by users on the GeoWiki platform.

proposed by the user input?	assumption is true or false. Retrieve all the objects of the selected owl species with the relation FoundOn	The different classes can be used to validate the assumption made by considering their exact geographic location	$(FoundOn)_i \rightarrow)$ to such species(S_i)]	$(HasLong)_k \rightarrow)$ $\rightarrow S_{ijk}]]$	"Owl" cs:HasLong ?o1 . "Owl" cs:HasLat ?o2 . } (Query only check for all forest habitat for owls in the dataset. Therefore, the results can be cross-referenced on the user input from the GeoWiki platform)
SET 2					
Q2.1[35] Where are most of the recently reported location of Waterfowl birds?	Retrieve all subject mapped to the values Waterfowls. Select the dates they were reported using the relation RecordedAt. Order in Descending other and limit to 10. This is to give the last 10 reported waterfowl species in time	For all subject (species) mapped with HasName and has the value waterfowls, select the Objects Values using the relation RecordedAt. Order by date and select the first 10. Label it with Y. Select the Locations of all Y using the relations HasLong and HasLat.	For all species in the dataset ($\forall$ S $\in$ N) select (|) [Subject (S_i) mapped with the relation HasName (←(HasName) →), and the Object value is Waterfowl (W)]. Save results as B (⋙ B). Select from result B the Species (S_i) and Objects (O_i) with relation RecordedAt (← (RecordedAt) →). Save results as Y (⋙ Y) and order in descending order with limit 10 DESC|10.	($\forall$ S $\in$ N) |[[S_i [S_i (←(HasName) →) →W] ⋙ B | O_i [O_i (← (RecordedAt) →)]]⋙ Y| O_{ij} [O_{ij} (← $(HasLong)_i \rightarrow) \wedge$ (← $(HasLat)_j \rightarrow)$ $\rightarrow S_{ij}$]] DESC|10	SELECT (DISTINCT ?s) (?o1 as ?lat) (?o2 as ?long) ?s WHERE{ ?s cs:HasName "Waterfowl" . ?s cs:HasLong ?o1 . ?s cs:RecordedAt ?o2 . ?s cs:HasLat ?o1 .

[35] Results not reported Yet

			Select the Objects(O_i) mapped to Subject (S_i) in Y with relations HasLat ($\leftarrow$(HasLat)$_I$ $\rightarrow$) and HasLong ($\leftarrow$ (HasLong)$_j$ $\rightarrow$)		} ORDER BY DESC|10
Q2.2 What are the characteristics of these locations?	Select the land use and land cover information of the Areas of Y.	For all triples with Y select the object that is mapped with the relation FoundOn.	Select Objects (Oi) mapped to Y using the relation FoundOn ($\leftarrow$ (FoundOn)$_i$ $\rightarrow$)	($\forall$ S $\in$ Y) | O_i [O_i ($\leftarrow$ (FoundOn)$_i$ $\rightarrow$) $\rightarrow S_i$]	SELECT (DISTINCT ?o as ?Land_Information) WHERE{ ?s ? cs:FoundOn ?o . }
Q2.3[36] What other species are available at those locations?	Select other species that are within the location of these areas. (Result from query Q2.1)	For all triples with Y select species using the relations HasLong and HasLat. Specie selection should be within a kilometer.	Select Objects ($O_{i,j}$) mapped to Y using the relations HasLat ($\leftarrow$(HasLat)i $\rightarrow$) and HasLong ($\leftarrow$ (HasLong)$_j$ $\rightarrow$)	($\forall$ S $\in$ Y) | O_{ij} [O_{ij} ($\leftarrow$ (HasLong)$_i$ $\rightarrow$) $\wedge$ ($\leftarrow$ (HasLat)$_j$ $\rightarrow$) $\rightarrow S_{ij}$]	SELECT (DISTINCT ?s) (?o1 as ?lat) (?o2 as ?long) ?s WHERE{ ?s cs:foundIn ?o . ?s cs:HasLong ?o1 . ?s cs:HasLat ?o2 .}
SET 3					

[36] The result of the Query does not end with the select clause. A detail analysis of all the species location should be checked with other GeoSPARL query

Q3.1[37] Which areas can support Forest?	Select all species in Alaska. Select Their longitude and latitude using the HasLong and HasLat Classes The results can be compared the Areas with high birds density. And the result can be used for making decisions on potential areas for developing vertical forest.	For all species in Alaska, select the objects values from the relation HasLong and HasLat. Cluster the species according to locations and select the most clustered locations.	For all species in the dataset ($\forall$ S $\in$ N) select (|) [Subject (S_i) mapped with the relation FoundIn ($\leftarrow$(FoundIn) $\rightarrow$) and the Object value is Alaska A] and saved as E ($\Rrightarrow$ E). Select the object values for Species in E with the relations HasLat ($\leftarrow$(HasLat)$_I$ $\rightarrow$) and HasLong ($\leftarrow$ (HasLong)$_j$ $\rightarrow$).	($\forall$ S $\in$ N) | S_i, O_i [S_i ($\leftarrow$(FoundIn) $\rightarrow$) $\rightarrow$A] $\Rrightarrow$ E | O_{ij} [O_{ij} ($\leftarrow$ (HasLong)$_i$ $\rightarrow$) $\wedge$ ($\leftarrow$ (HasLat)$_j$ $\rightarrow$) $\rightarrow$ S_{ij}]	SELECT (DISTINCT ?s)(?o1 as ?Longitude) (?o2 as ?Latitude) (COUNT(?S) AS Number ?) WHERE{ ?s cs:foundIn "Alaska" . ?s cs:HasLong ?o1 . ?s cs:HasLat ?o2 . } Another Version that can produce similar result by considering the proposed classes in Table 8-2 SELECT (DISTINCT ?s)(?o1 as ?Longitude) (?o2 as ?Latitude) (COUNT(?S) AS Number ?) WHERE { ?s1 cs:Breeds "Dragonfly" . "Dragonfly" cs:FoundOn ?o4

[37] The result of the Query does not end with the select clause. A detail analysis of all the species location compare to acceptable areas must be done by the users of the information.

					?s cs:Consume “Necta” . ?s cs:HasLong ?o1 . ?s cs:HasLat ?o2 .} Note[38]

Note the Math relation is in the form Select (object or subject) From (Types of relations) Where (Relations under consideration). Most often the from clause is all the triples in the dataset under considerations

Table 0-5: Tool Comparison for Data Integration: Source: Author

	Karma Data Integration Tool	COMA CE	The Talend data Studio
Quality Criteria			
Support	There exists direct support for the tool because the project it was developed for is still ongoing. There are available source codes that can be edited by users to suit the needs of one’s job requirement.	There is no link to direct support. However, the platform is active, and there is high-level versioning on GitHub. The tool is designed for ontology mapping, and semantics and active community work on the tool.	Direct support available. However, the tool needs a license. The license can be obtained at a cost
Searching Capabilities	It provides a searcher for looking for the most appropriate definitions and cross-referencing with other existing ontologies.	The tool does not provide a searcher section for looking for the most appropriate definitions and cross-referencing with other existing ontologies	A searcher for looking for the most appropriate definitions and cross-referencing with other existing ontologies
Analyzer	The tool provides means of structuring the lexical and syntactic semantics by providing a	The tool does not provide any analyser for checking lexical and syntactic errors.	The tool provides means of structuring the lexical and syntactic semantics by

[38] The different relations are meant to enrich the selection process.

	layer that serves as an analyser to guarantee the absence of lexical and syntactic errors.		providing a layer that serves as an analyser to guarantee the absence of lexical and syntactic errors.
Usage	The tool accepts different file formats which can be modelled according to the generated ontology. Pub[39]	The output file is only the R2RML mappings which can be used in another platform with the datasets	The tool accepts different file formats which can be modelled according to the generated ontology. Pub[40]
Technicalities	Requires a localhost server for maintenance and updating User impute. The localhost server provides means of maintaining and managing the information provided	Provides a server which is hosted which is not locally hosted. Therefore, the client has to pay for the services.	Provides a server which is hosted which is not locally hosted. Therefore, the client has to pay for the services.
Capabilities	The tool can import different dataset with different data format during the integration process. Moreover, the tool can support different ontologies at the same time. Therefore, different works can be performed concurrently. The Mapping and merging process uses based on the semantics of both the datasets and the ontology.	The tool takes only one ontology and one dataset at a time. Mapping is based on standard vocabulary among dataset and ontology. Therefore, each dataset can be modelled base on the ontology schema and compared with different datasets in another round of modelling.	Fewer capabilities. However, the tool has a defined scope.
Efficiency	Efficient[41]	Efficient.	Fast in computation and easy to analyse large sum of datasets

[39] The output file can be published in different file formats for easy usage.
[40] The output file can be published in different file formats for easy usage.
[41] The tool exhibits high ability to integrate huge datasets efficiently. There exists a function that gives the possibility to alter the speed and memory usage.

Documentation	Proper Documentations Available[42]	Proper Documentations Available[43]	Proper Documentations Available[44]
Ability to debug	The platform is currently active. Therefore, debugging issues can be resolved in time.	Less debugging notes and platform available	Less usage and no debugging records found.
Modifications ability	The tool does not provide an editor for adding, modifying and removing definitions.	The tool does not provide an editor for adding, modifying and removing definitions in the ontology.	The tool provides an editor for adding, modifying and removing definitions.
Explanation for some of the criteria			
Support	Available support considers the availability of user-friendly assistance on technical problems. There should be available personnel (technical team) to help solve technical issues. The technical team should include professionals with adequate technical Know-how. Moreover, the technical support should be reached with an appropriate time.		
Capabilities and Richness	The tool should be capable of integrating different datasets with the different data format. There should be the possibility to integrate datasets and formats such as GEOJSON, XML and others concurrently. The integration should be based on the semantics and schema of the designed ontology.		
Efficiency (Speed)	The rate of processing different datasets should be fast and quick enough to process different datasets at a given time. The tool should have an adequate processing capability.		
Documentations	There should be adequate documentation on the usage and configuration that can enable users to use the tool efficiently. Proper documentation on a system allows flexible and adequate usage. Adequate documentation on a system plays an essential role in communicating the purpose of the application to users (Nasution & Weistroffer, 2009).		
Bugs and debugging issues	There should be some individuals (Users) using the application with potential information on buds and debugging issues. The more people use a particular piece of software, the more errors associated with the application can be realised. Therefore, a potential criterion for selecting an application is the number of users (popularity)		

[42] Well -documented. The tool provides well-structured documentation on usage and definition of all component
[43] Well -documented. The tool provides well-structured documentation on usage and definition of all component
[44] Well -documented. The tool provides well-structured documentation on usage and definition of all component