# Design and ImplemEntation Of The Ontology

This chapter describes the design and implementation of the geo-citizen science ontology**.** It discusses the granularity, formalisation 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.

## Introduction:

The process of building the citizen science ontology is based on the IEEE software development life cycle (IEEE, 1997). The framework is a standard that provides clear and precise structure for building software applications. The overall steps to be followed is shown in Figure 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 1 shows the general overview of the design of the ontology using the two fused frameworks.

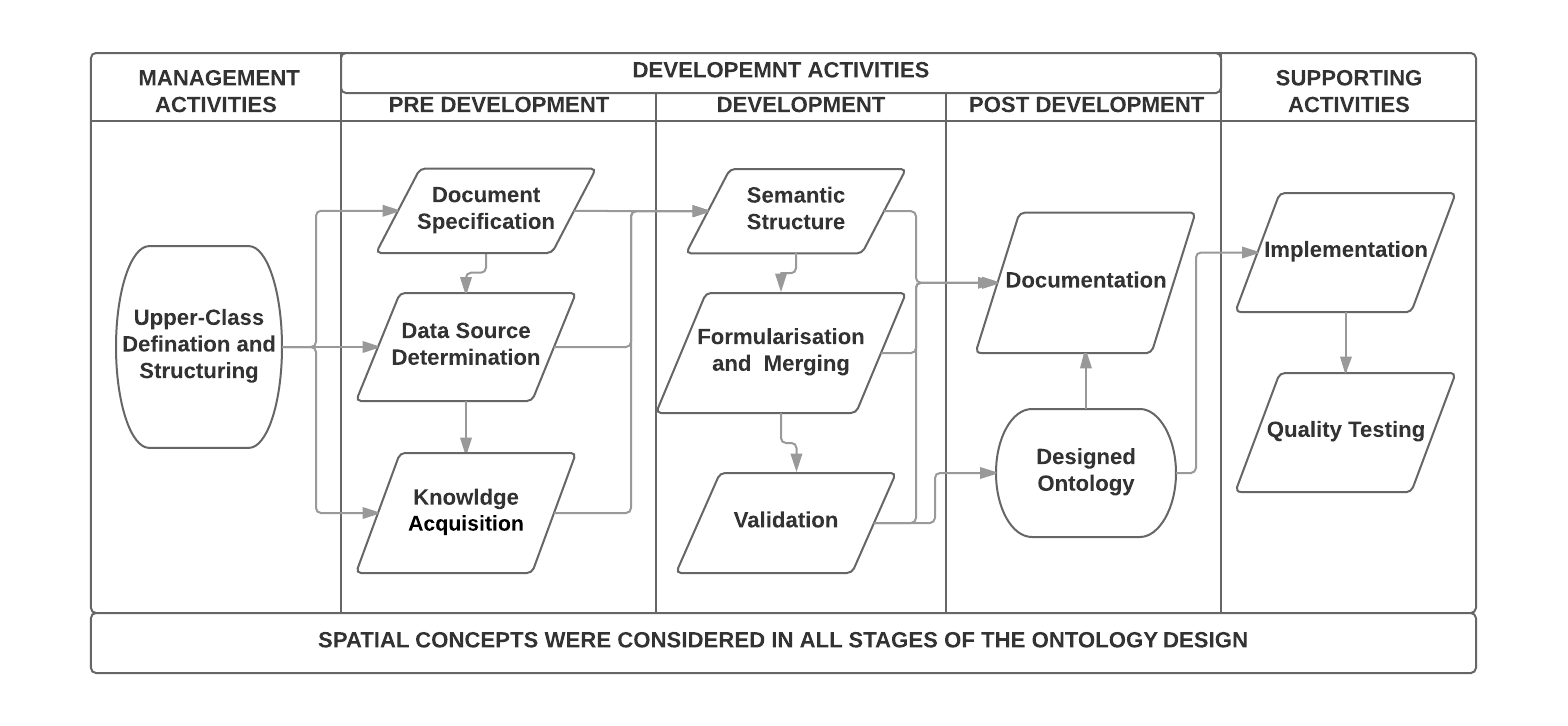


Figure 1: Overall Structure for the Ontology Design and Implementation

## Ontology Management

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 grouping of ontology classes based on a higher-level abstraction to accommodate the different part of citizen science domain.

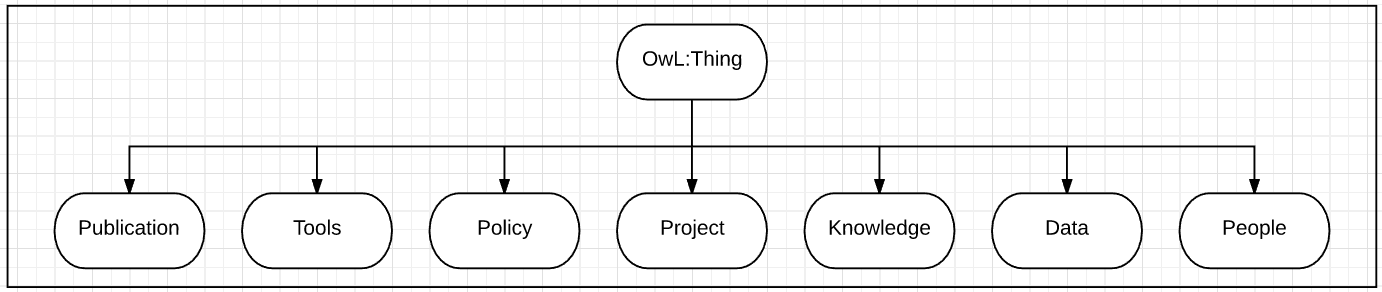
The Upper-Level ontology classes in the ontology design serve as a level of abstraction, which gives accurate categorisation to the ontology to accommodate different sections in the ontology design for different purposes. However, these levels of abstractions are not referred to any identifiable, concrete 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 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 2: Upper-Level Ontological Classes

## Development (Generic Ontology Development Framework)

The ontology development sections report on the design of the geo-citizen science ontology using the Generic Ontology Development Framework at the development section for the adapted IEEE framework. Figure 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)

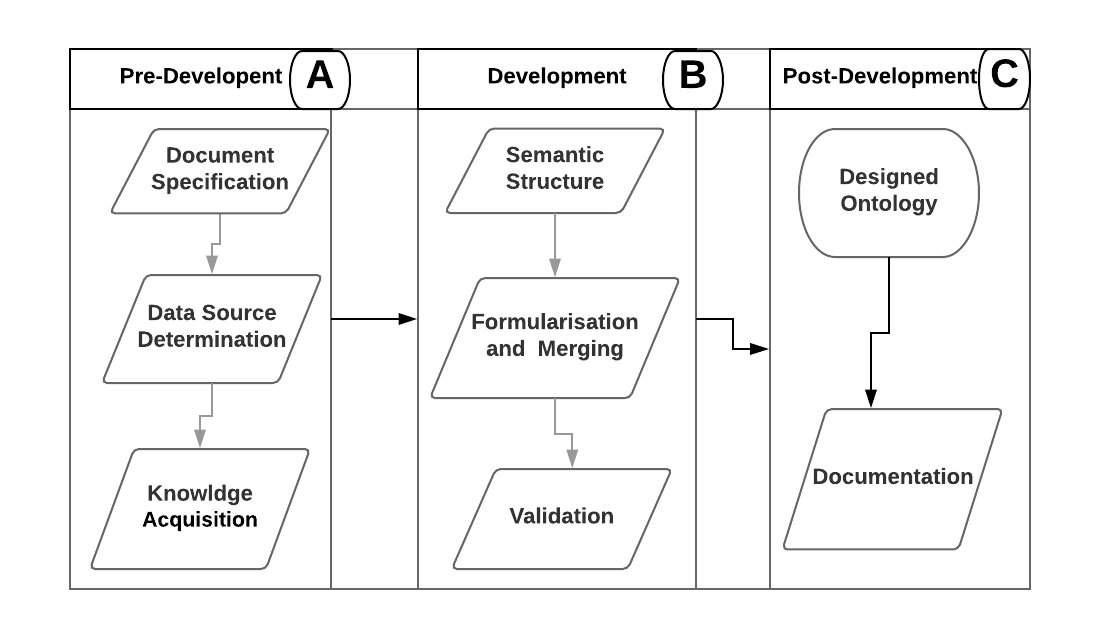


Figure 3: The Designed Framework for the Ontology Development Section

### 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 geo-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 4 shows steps followed to acquire all the needed information in the development stage. Form Figure 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.

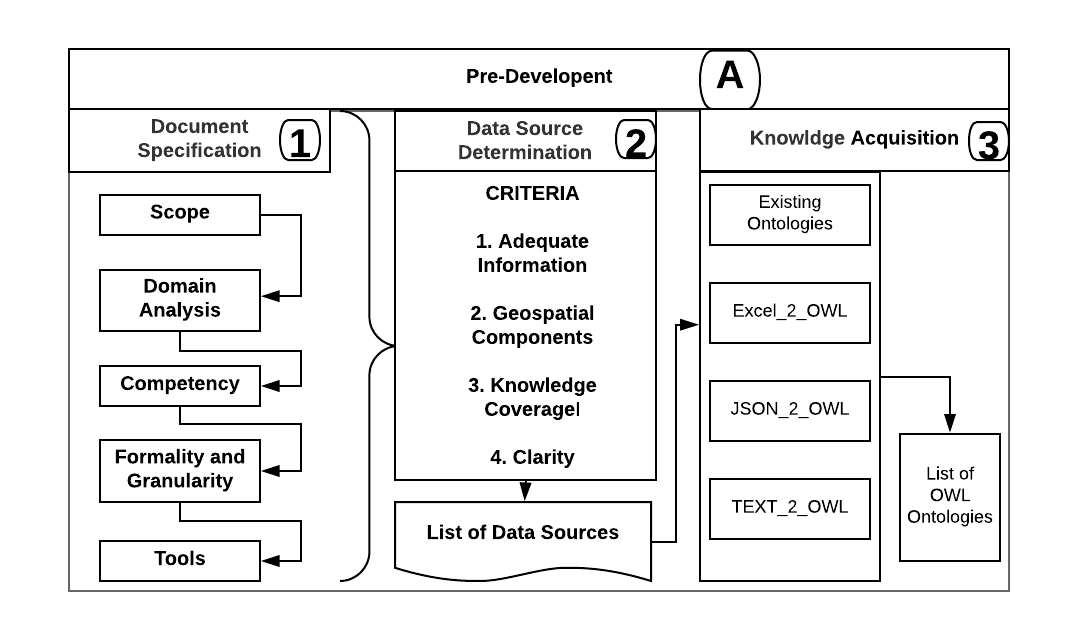


Figure 4: Overall Steps at the Pre-Development Section

**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 envisage 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 envisage ontology will consider the domain of environmental and biodiversity. Table 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 geo-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.

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.

1. **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.
2. **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.

1. **Structure and formats of the datasets:** The primary purpose of the geo-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 1: Selected projects to be considered in the knowledge acquisition stage.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Project Name | Domain | Link |
| 1 | Geowiki | Land Resources (maps) | [Link](https://www.geo-wiki.org/) |
| 2 | Did You Feel It? (DYFI) | Earthquakes | [Link](https://earthquake.usgs.gov/data/dyfi/) |
| 3 | Atlas of Australian Birds | Birds | [Link](http://birdlife.org.au/projects/atlas-and-birdata) |
| 4 | Anecdata | Botany, Entomology, water quality, Phenology, air and water quality etc. | [Link](https://www.anecdata.org/register) |
| 5 | Big Bug Hunt | Entomology | [Link](http://bigbughunt.com/) |
| 6 | Big Butterfly count | lepidopterology | [Link](http://www.bigbutterflycount.org/) |
| 7 | GeoTag-X | Disaster Risk Reduction | [Link](https://geotagx.org/) |
| 8 | iNaturalist | Biodiversity | [Link](https://www.inaturalist.org/) |
| 9 | NatureWatch | Ice, Frogs etc. | [Link](https://www.naturewatch.ca/download/) |
| 10 | Mangrove Watch | Wetlands Monitoring | [Link](http://www.mangrovewatch.org.au/index.php?option=com_wrapper&view=wrapper&Itemid=300390) |

**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 geo-citizen science ontology. Table 2 shows a general review of the groupings from the citizen science projects. The review in Table 6 explains the domain of interest and the intended purpose of the final geo-citizen science ontology. The diversity of the domain of citizen science makes it more exciting and time-consuming to model the geo-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 potential frameworks to be used are discussed in chapter four.

Table 2 Granularity and Formulation of the Ontology Under the Upper-Class Knowledge for both Domain-Specific and General Domain Concepts

|  |  |
| --- | --- |
| 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, there is numerous citizen science projects that 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, 1997). Therefore, there is numerous citizen science projects that collect information on plants for research activities. Examples of these projects are the [NatureWatch](https://www.naturewatch.ca/download/) from Canada and the [Inaturalist](https://www.inaturalist.org/) 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](https://www.amentsoc.org/insects/glossary/terms/biological-classification)) | Classification of species into groups due the common characteristic possessed by such species helps in identifying unique 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 geo-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 geo-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 7 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 3 gives an overview of the datasets. The following conditions were observed before formulating the Table.

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 comes 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 3 Overview of Datasets

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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-structured information in the form of metadata. The format selected for this datasets is Excel (CSV). | ([link](https://earthquake.usgs.gov/earthquakes/search/#%7B%22feed%22%3A%221437493916387%22%2C%22search%22%3A%7B%22id%22%3A%221437493916387%22%2C%22name%22%3A%22Search%20Results%22%2C%22isSearch%22%3Atrue%2C%22params%22%3A%7B%22producttype%22%3A%22dyfi%22%2C%22orderby%22%3A%22time%22%7D%7D%2C%22listFormat%22%3)) |
| 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](https://www.geo-wiki.org/downloads/)) |
| 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 is in four different categories. Frogs, Plants, Ice, Worm and Milkweeds. | ([link](https://www.naturewatch.ca/download/)) |
| 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](http://www.bigbutterflycount.org/2017mainresults)) |
| 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](https://www.anecdata.org/posts)) |

#### 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 are explained in the appropriate section.

**Reuse of Existing Ontologies: *(Existing Ontologies*)** The following ontologies are evaluated to be used in the geo-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 4 shows some selected existing ontologies that were reviewed and used in the geo-citizen science ontology.

Table 4: Selected Classes from Existing Ontologies Reuse component

|  |  |  |
| --- | --- | --- |
| Ontology | Number of classes | Area of interest |
| Plant ontology | 30 | Plant descriptions with plant types and plant names |
| Vertebrate Taxonomy Ontology ([link](http://www.obofoundry.org/ontology/vto.html)) | 25 | Vertebrate descriptions with plant types and plant 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 ontologies | All | Spatial relations to map spatial entities |
| W3C Geo vocabulary |  |  |

**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 base on the semantics of the data and the datatypes in the data. Figure 6 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 6, 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. As displayed in Figure 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](https://github.com/protegeproject/cellfie-plugin) points to the source of the plugin.

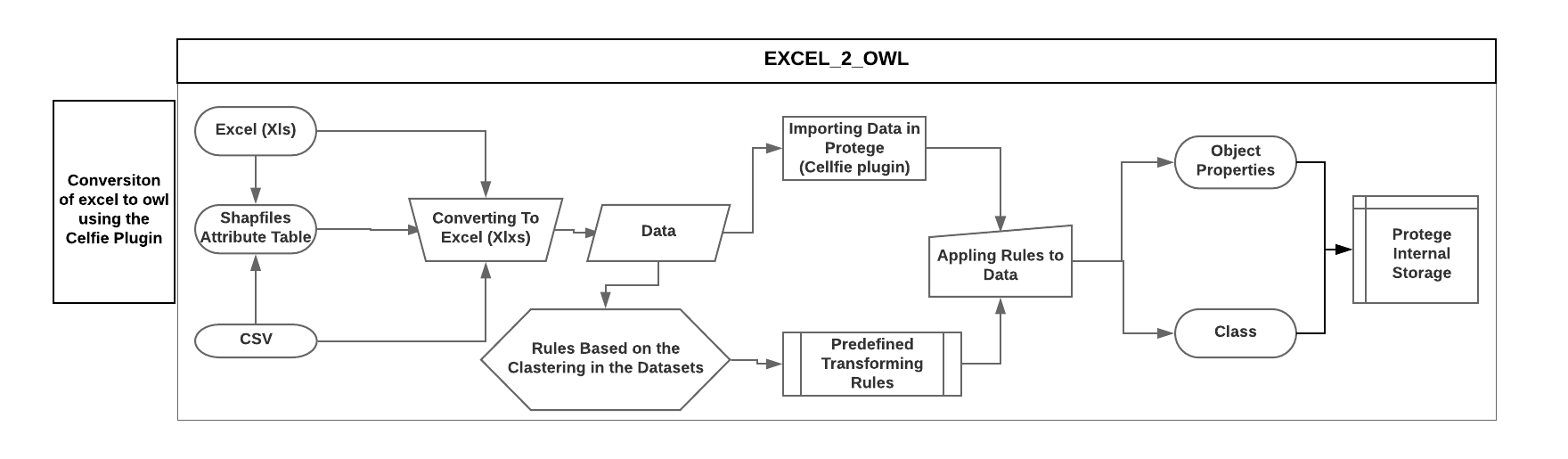


Figure 5: Overall Conversion Process from Excel to OWL Ontology using the Cellfie Plugin

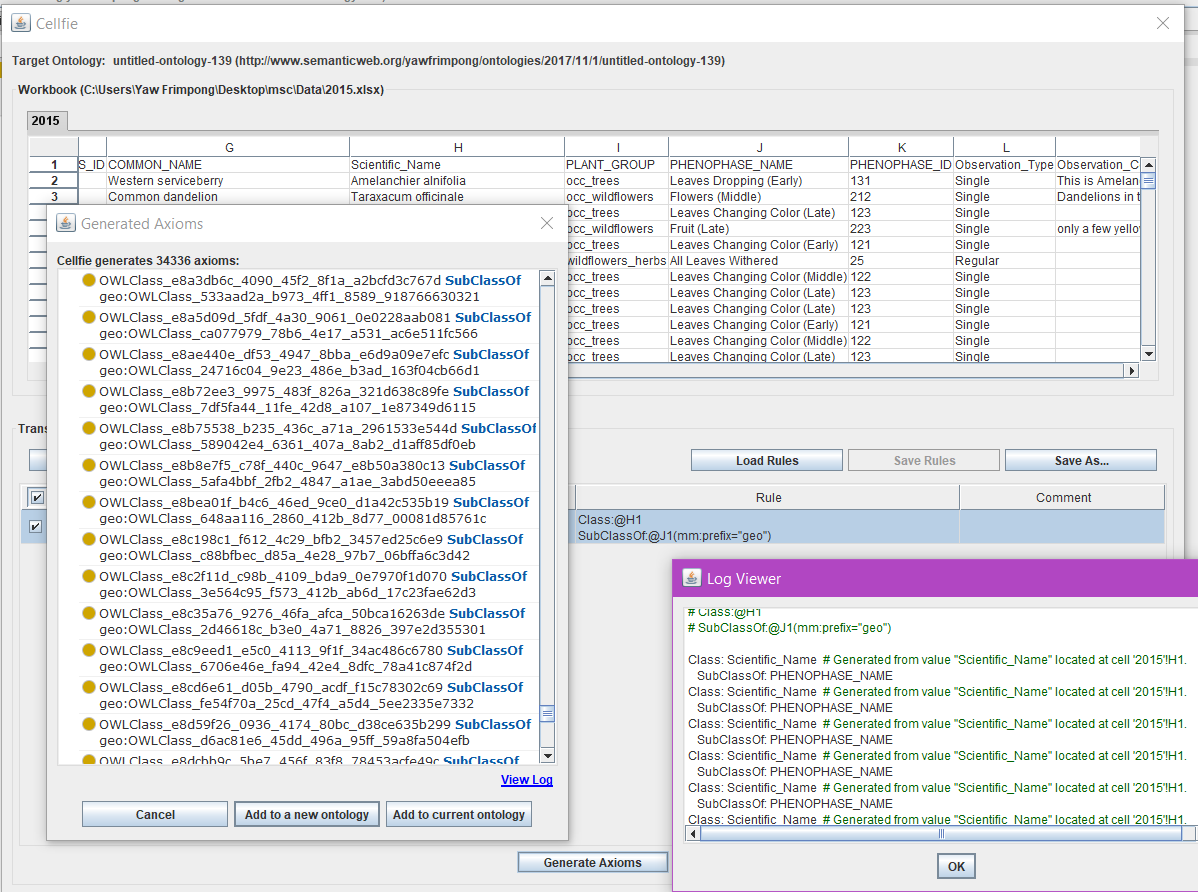


Figure 6: Data Imported and Mapped with the Cellfie Plugin in Protégé

**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 7 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. Figure 4 shows the interface of the tool after importing data and parsing to OWL ontology. The tool was obtained from this [link](https://github.com/2529742/json2owl) and modified by the author.

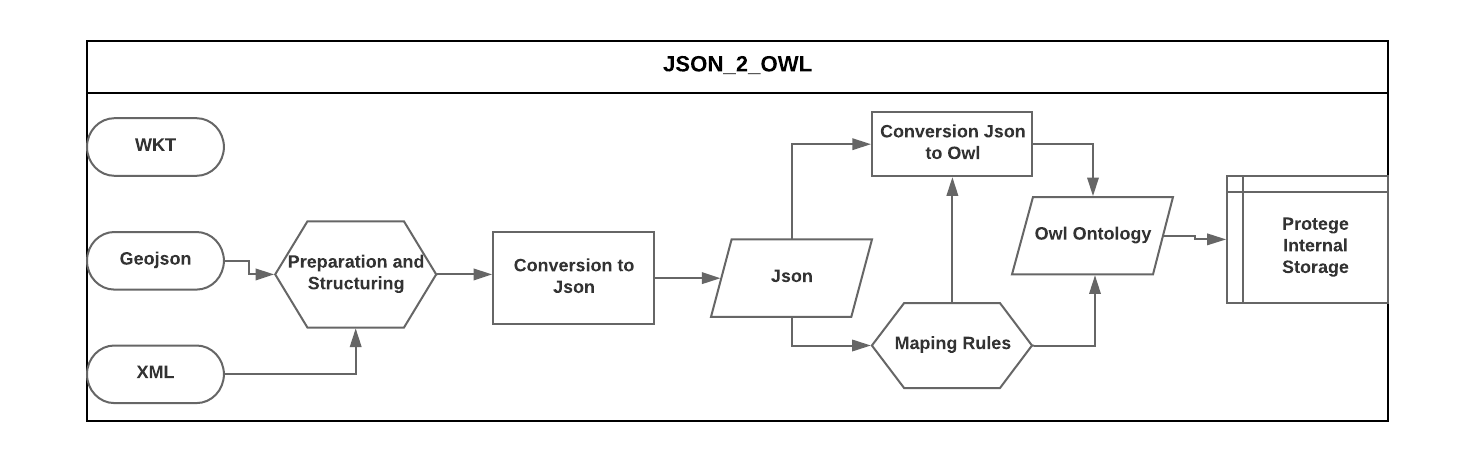


Figure 7: Overall Process for Converting JSON to OWL in the Ontology

**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 8 shows the workflow for the conversion from text to OWL.

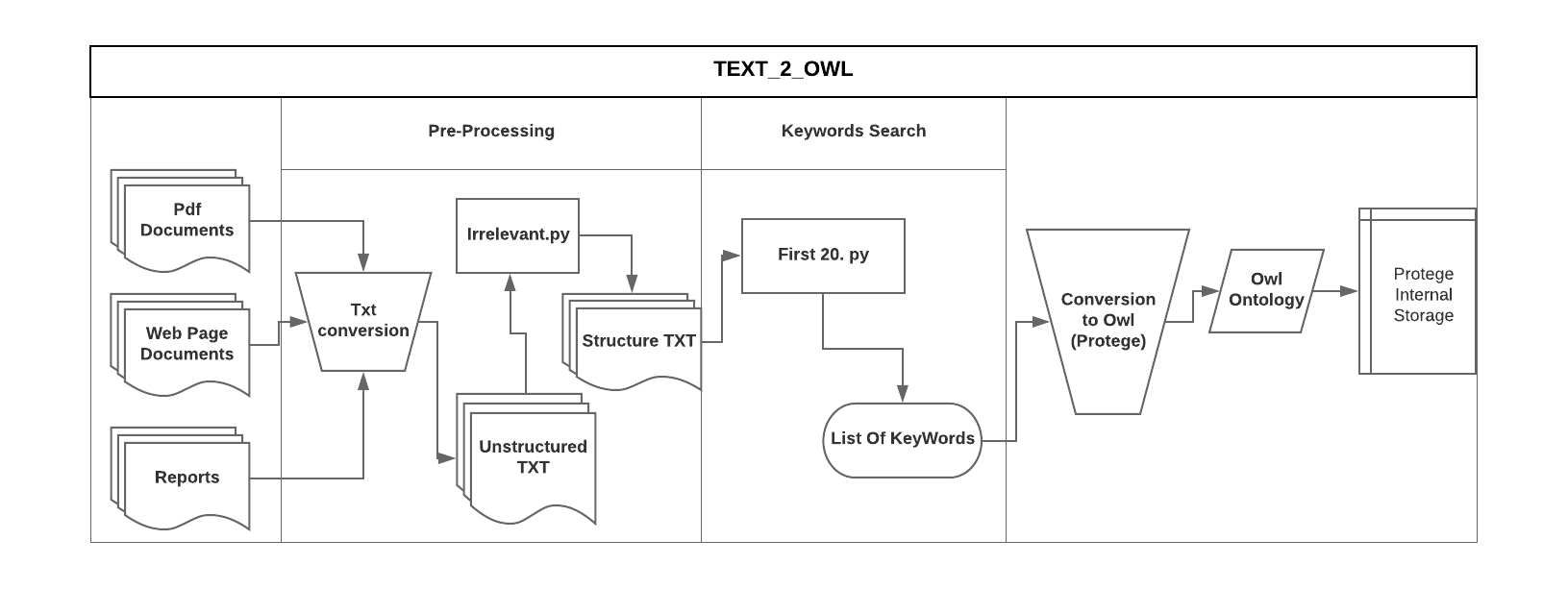


Figure 8: Steps in Converting Text File to Owl Ontology

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

### Development Phases (Ontology Development) (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 1 shows a general overview of the steps considered at this stage.

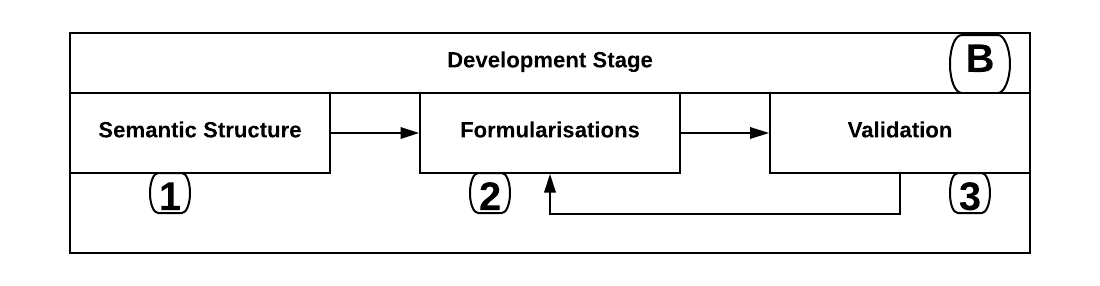


Figure 9: General Overview of the Development Section (Generic Ontology Development Framework)

The upper-level class **Knowledge** has 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 is used in the formalisation 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 key elements of knowledge capturing in the ontology design. The semantic structure serves as a way of ensuring a logical flow of the concept in the 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

Oi = {Ci, CisSubClass, RelCi→cj Instanceci Rulesoi, Axiomsoi}

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, (Oi):** 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 geo-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 (Dean & Owl, 2004). 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 is obtained from the Plant Ontology (PO) analysed for reuse. Due to the granularity and scope of the geo-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. The table 10 in appendix I gives some examples of classes used in the geo-citizen science ontology.

**Subclass, (CisSubClass):** 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 inherent in it subclass. 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 (RelCi→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. Examples of the relations used in the ontology can be found in the Appendix.

**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 (Rulesoi):** 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 (Axiomsoi):** Class axioms help express class descriptions which form the basic 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 subclass of the superclass nature defined as a concept for both domain specific and generic concepts in the ontology. Class prefix are omitted for clarity and simplicity. Words in the form **Grass** representclasses, and ***HasType*** represent Relations or data properties

In the biodiversity domain, more classes were initiated. However, considering the integration of biodiversity conservation into city conservation planning, two top-level classes **Nature** and **Concepts** were realised. The purpose of the superclass **Nature** is to capture the knowledge in the domain of both biodiversity and landscape models which includes land use and land cover information. Whiles Concepts is to define a conceptual framework for modelling data during the data integration process.

The class land is a subclass of Nature which has a subclasses **LandUse** and **LandCover**. **LandCover** **HasTypes** **Grass**, **Asphalt**, **Trees**, **BareGround**, **Water**. However, **Grass** falls under class **Plant** and has ***SubClasses*** **Wheat**, **Oat**, **Teff** and others. **LandUse** also ***HasTypes*** **Recreational** (**Parks** ***Is-A-Type-Of* Recreation**), **Transport** (**Roads**), **Agriculture** (**Farmlands**), **Residential** (**Housing**), **Commercial** etc. **Agriculture** includes **Grass**. 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.

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**. 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). The class **Animal** expresses knowledge of different animals. It however considered potential classes from both the BBC Wildlife Ontology and the Geospecies Ontology. A subclass is the class **Insects**. Class **Insect** expresses the types, social behaviour and different examples of insects. Most of these knowledge is obtained from resued classes from the Social Insect Behavior ontology (SIBO).

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.

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. 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. This relation was reference and assigned with the same as 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 what a location of an object is, 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 are considered to be within a city’s boundaries. A **City** has both **LandUse** and **LandCover** that determines the required conservation plan.

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**) 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.

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 is aligned and merge 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**. A second example is Herbivore, carnivore, and omnivore that were mapped as subclasses of feeding in the class animals.

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. The triple ‘**Waterfowls** ***eat*** **Insects’** such as **Dragonfly** was was instantiated. Other mappings generated are **DragonflyNymphs** ***GrowOn*** **Wetlands**. **Waterfowl** ***FoundOn*** **Wetlands**. Insect lay their eggs on **WaterlogsAreas** which give their eggs nurturing condition for proper growth. Therefore, the class Insect and birds in the ontology are mapped with the relation eat. However, under the Class Insect in the ontology, there is the subclass **DevelopmentStage** which ***HasType*** **Nymph** as a ***SubClass***. ‘**Nymph** ***GrowsOn*** **Wetlands’** serves as a data property in the ontology for the class **Nymph**.

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. **Wetland** ***SubClassOf*** **LandCover** Inhabits **Waterfowl** ***SubclassOf*** **Birds** (N-rry triple formulation). Therefore, multiple sightings of the waterfowl birds in a particular location can be said to be a Wetland. **ShallowWateryAreas** ***Is-A-SubClass-Of* Wetlands** which includes **Swamps.** This is mapped with the relation **Swamp** ***Is-A-Type-Of*** **ShallowWateryAreas**. Another relation is **Owl** ***FoundIn*** **Forest**: This relation provides the notion that most of the owl sightings are found in forest 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.

**Swamp *Is-A-Type-Of* ShallowWateryAreas** and **Wetlands** ***Is-A-Type-Of* ShallowWateryAreas**: The two subclasses of shallow watery areas can aid in validating Wetland land class.

**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](https://a-z-animals.com/reference/glossary/#jump-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. **Grass** ***SubClassOf*** **Plant** ***FoundOn*** **Grasslands** are possible integration strategies adapted during this formularisation stage.

The Use Case “Developing Vertical Biodiversity”, vertical forest can exist at different height levels. Information on different animals for specific height was handy. Under the class, **Domain** has ***SubClass* Ecology.** **Ecology** has ***SubClass*** **Food** which describes the different species and their diet. 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 geo-citizen science ontology captures most of the knowledge required 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, and contradictory concepts were resolved manually. The final consistent and uncontradicted ontology was converted into owl dl. The formalisation under the protégé environment allows conversion of the generated ontology to different formats 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 meaning contained in the two classes.

### Post Development Stage (C)

### 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 intended 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 full documentation of classes and their intended means is shown in Appendix 1 with the glossary of terms.

**Designed Ontology**

The final design ontology is made of list of concepts in the domain of citizen science internally stored in the developing environment of protégé. Figure 10 shows a preview of the developed ontology in the protégé environment.

Individuals to be generated in the ontology are to be mapped to other individuals to ensure compatibility with each other. However, the mappings among individuals were conducted based on the relations that exist among classes. Figure 11 and 12 show the general overview of how ontological classes and relations are mapped both among classes and individuals in different classes. 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 classes in the ontology than the normal instances of the class. 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 2, 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*. Individuals in a class are considered to have inherent relations with that class. From figure 1, 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 each other by considering the datatypes for each data property relevant for the specific class. An example is the relation 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. The table 1 below aims at presenting each competency question in the use case and review possible strategies for connecting individuals in the datasets during the mapping and modelling stage.

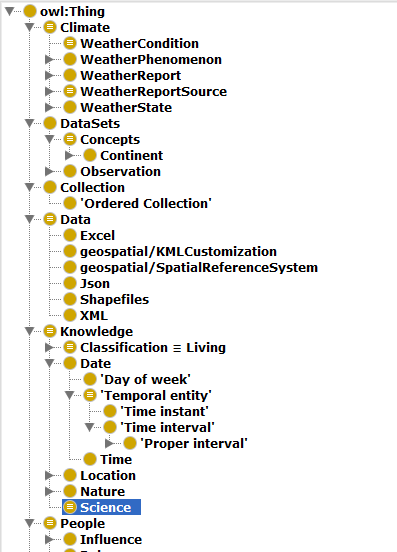
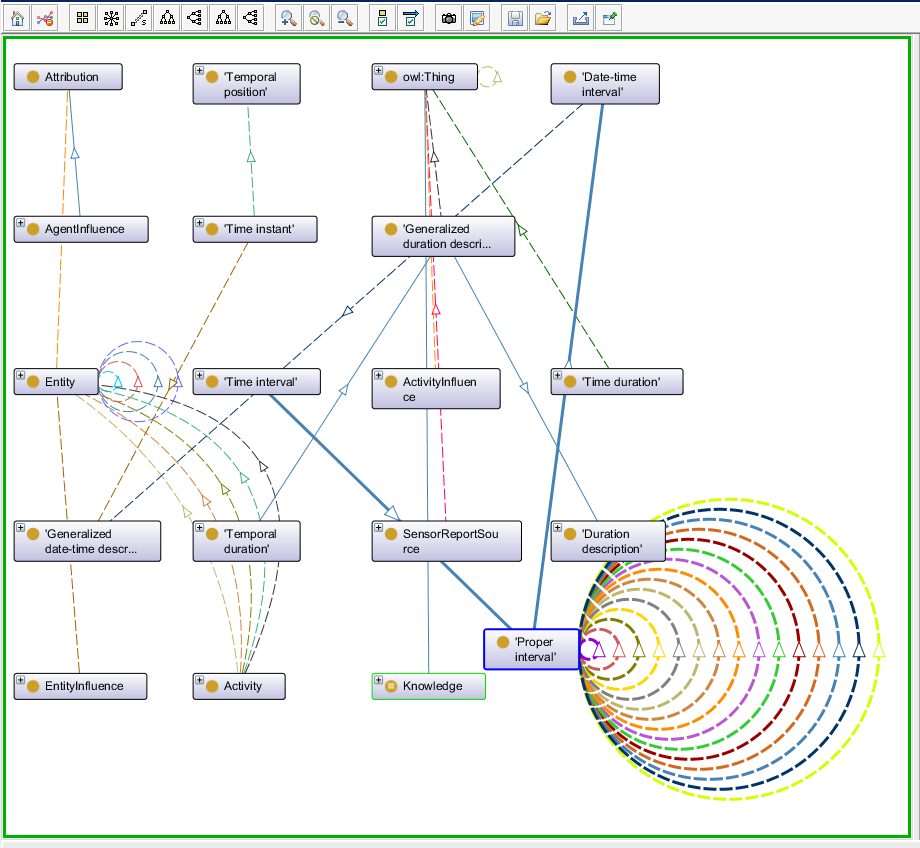


Figure 10: Preview of the Designed Ontology

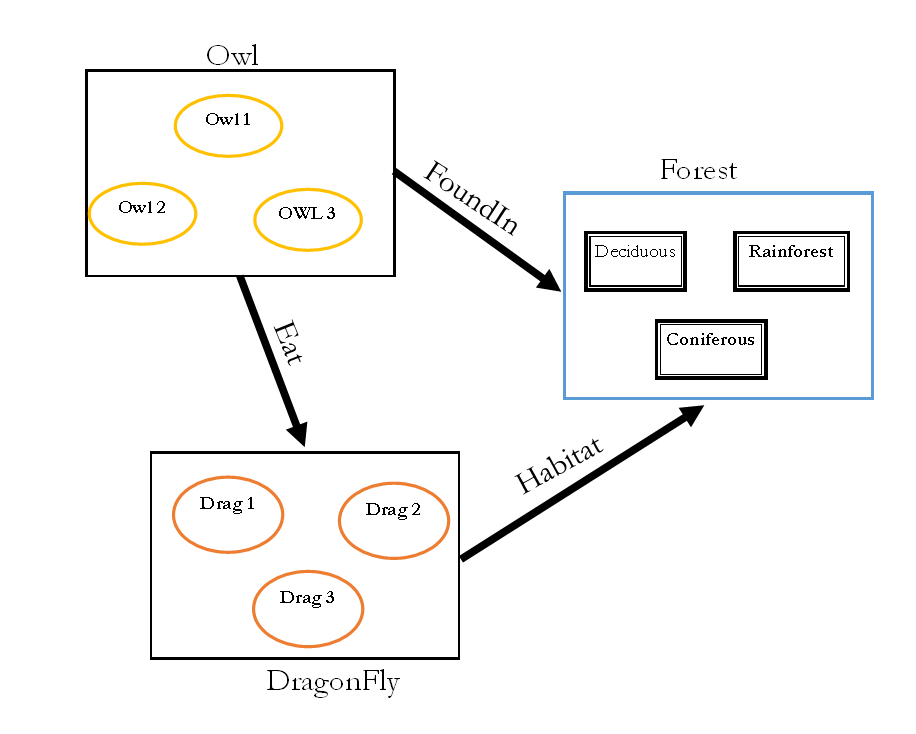


Figure 11: Mapping Ontological Classes

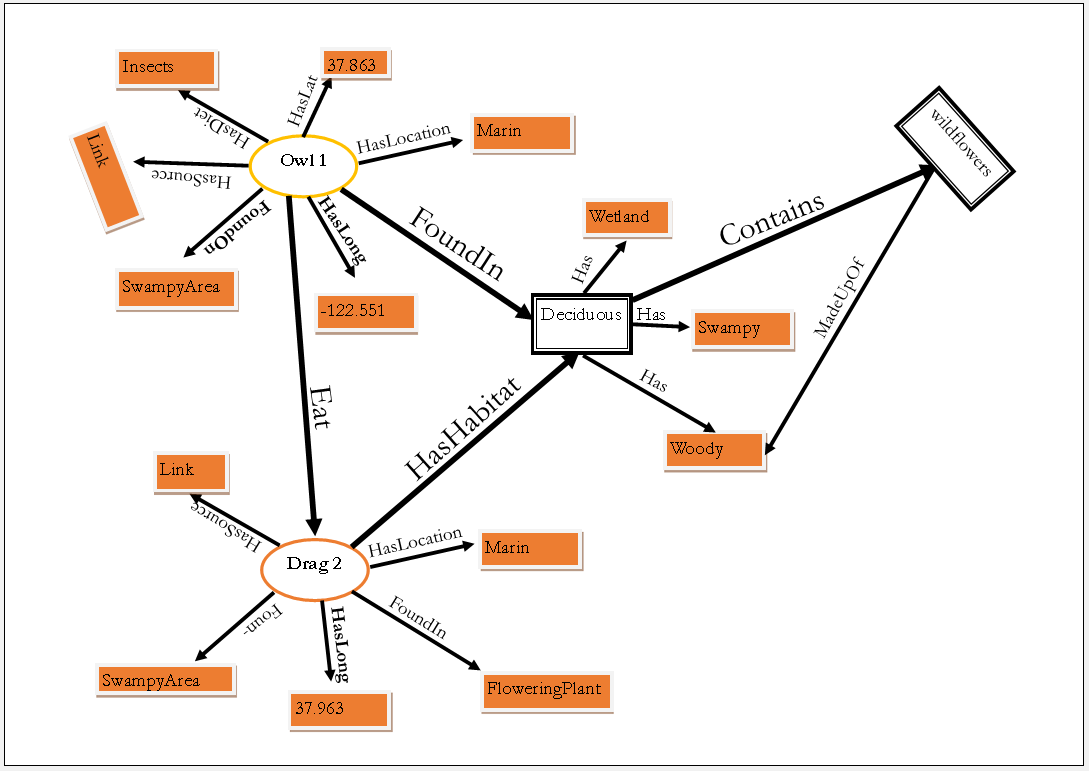


Figure 12: Mapping of Individuals Using both Object and Data Properties

## Ontology Supporting Activities

This section gives an overview of the implementation strategy and a general overview of the quality testing for the designed geo-citizen science ontology. Details of the quality testing can be found at chapter 6

### Implementation

The task of implementing the ontology in an ontology language requires an environment that supports the ontologies selected in the integration step. Features that should be provided by such an environment are

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 during the

The process described in Section 4.3.2 is now formalised into the formal model of the ontology language being used. In the case of the Geo-citizen science ontology, an OWL Ontology is created using Protégé together with the Pellet reasoner.

After exhaustive analysis and structuring in the previous sections, the step of implementing the ontology has become a straight-forward task. The geo-citizen science ontology is implemented in OWL using Protégé 4.3 together with the Helmet OWL 2 Reasoner. Helmet is an ontology performance tool that uses a set of patterns to find possible performance problems in an OWL ontology. Helmet has been used intensively to ensure it does not report any problems that could affect reasoning performance

### Quality Testing Strategies

The quality of the generated ontology is assessed based on the use of the ontology for the integration of data ontology in the citizen science domain. A second quality testing strategy used I this section is the Semiotic metric suite by Burton et al., (2005). The integration strategy with data is composed of a review of different implementation tools. Examples of such tools are the Karma Data Integration tool and the Coma Ce 3.0 tool. The metric suite quality testing be assessed in the following ways: maintainability, understandably, adoption, and application performance