# The Role of Semantic Web Technologies for IoT Data in Underpinning Environmental Science


**Izhar Ullah**

**This dissertation is submitted for the degree of Doctor of Philosophy**

**October 2018**

**School of Computing and Communications**

*To my late father …*

# Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisor Prof. Gordon Blair.

Izhar Ullah, MSc

Lancaster University, UK

# Abstract

The advent of Internet of Things (IoT) technology has the potential to generate a huge amount of heterogeneous data at different geographical locations and with various temporal resolutions in environmental science. In many other areas of IoT deployment, volume and velocity dominate, however in environmental science, the more general pattern is quite distinct and often variety dominates. There exists a large number of small, heterogeneous and potentially complex datasets and the key challenge is to understand the interdependencies between these disparate datasets representing different environmental facets. These characteristics pose several data challenges including data interpretation, interoperability and integration, to name but a few, and there is a pressing need to address these challenges. The author postulates that Semantic Web technologies and associated techniques have the potential to address the aforementioned data challenges and support environmental science. The main goal of this thesis is to examine the potential role of Semantic Web technologies in making sense of such complex and heterogeneous environmental data in all its complexity.

The thesis explores the state-of-the-art in the use of such technologies in the context of environmental science. After an in-depth assessment of related work, the thesis further examined the characteristics of environmental data through semi-structured interviews with leading experts. Through this, three key research challenges emerge: discovering interdependencies between disparate datasets, geospatial data integration and reasoning, and data heterogeneity. In response to these challenges, an ontology was developed that semantically enriches all sensor measurements stemmed from an experimental Environmental IoT infrastructure. The resultant ontology was evaluated through three real-world use-cases derived from the interviews. This led to a number of major contributions from this work including: the development of an ontology tailored for streaming environmental data offering semantic enrichment of IoT data, support for spatio-temporal data integration and reasoning, and the analysis of unique characteristics of environmental science around data.

# Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Gordon Blair for his invaluable guidance, and substantive support in the realisation of this work. Gordon has been a great mentor to me throughout these four years. Without his support, this thesis would have never been completed. To be honest, I am very lucky being his doctoral student. I would also like to thank Dr. Vatsala Nundloll for her support and crucial feedback on all important milestones in the PhD process. She had been always there when I needed to discuss. I would like to thank Dr. Suhail Yousaf and Dr. Yusuf Sani for their guidance, critical analysis and moral support throughout my PhD. I owe a great debt of gratitude to all those environmental scientists who participated enthusiastically in my semi-structured interviews and provided me a highly rich set of qualitative data and knowledge about their domain. I would like to thank all my colleagues in the School of Computing and Communications who made this journey less lonely and more enjoyable, particularly Faiza Samreen, Dr. Abdessalam, and Roberto for the fruitful discussions and the enriching feedback on my work. I would like to give a special thanks to my wife, mother, brothers and sisters for their unconditional love and support. Finally, I would like to thank the Commonwealth Scholarship Commission, London, UK, for granting me such a high-prestigious award to undertake my doctoral research. This thesis would not have been possible without their financial support.

Thank you all,

Izhar Ullah

# Contents

# List of Tables

# List of figures

# 1 Introduction

The advent of advanced computer and information technologies has changed almost every scientific and engineering field, introducing new ways of research based on data which has converted many disciplines from “data-poor” to “data-rich” environments [1]. There is a spectrum of how data underpins contemporary science. At one end of this spectrum, usually termed as the head, lies big science or data-intensive science, in which survey satellites, modern telescopes, high-throughput instruments, sensor networks, accelerators and supercomputers have been generating enormous amount of data in various disciplines like High Energy Physics, Astronomy, Life Sciences, just to name but a few [2]. These datasets, usually held by a few custodians, are: very large in size, most likely homogeneous collections with standard data format and uniform procedures, receive proper curation and maintenance, provide open access and reused effectively [3]. In contrast, the other end of the spectrum is commonly termed as the long tail of science which contains a large number of potentially small and heterogeneous collections of datasets [3]. These datasets are usually collected by individual scientists, small laboratories and/or projects. When combined together, they form a big portion of the data spectrum.

The long tail data exists in many sciences and environmental science is one such good example. Environmental science is an integrative, interdisciplinary and collaborative discipline which entails interaction between the four segments of environment, i.e. atmosphere, hydrosphere, lithosphere and biosphere [4]. It encompasses various sub-disciplines like biology, ecology, ethology, hydrology, soil science, biogeochemistry,

climatology, meteorology, oceanography and geography. Environmental scientists have been facing complex and unique challenges pertinent to human society, for instance, modelling future climate change scenarios, considering impacts of extreme events and maintenance of biodiversity [5]. However, the recent advancement in information science and technology in general and the environmental sensors and the Internet of Things (IoT) technology in particular has shaped environmental science considerably [6]. These contemporary technologies, providing real-time spatio-temporal data, have been playing a key role in understanding and managing the aforementioned environmental issues [7]. These in situ sensors, usually part of a wireless sensor network, monitor different environmental facets in the environment and generate enormous amount of data. On the one hand, there lies significant value in this data by enabling the discovery of hidden patterns in it. On the other hand, this data deluge leads to computational and statistical problems [8]. The analysis of this data (via data science) is distinct in environmental science, with its own particular challenges. In the data science literature, the three 'V's are often discussed, i.e. volume (the size of the datasets), velocity (the rate at which the data is generated) and variety (the range and heterogeneity of data sources). In many areas, volume and velocity dominate and computer scientists face the challenge of efficient processing of potentially massive datasets. But in environmental science, the more general pattern is quite different and often variety dominates. This equates to the long tail of science introduced above where data is obtained from diverse data sources with different data formats, at different geographical locations, and with various temporal resolutions. These key features pose several data challenges including data interpretation, interoperability and integration, to name but a few. These challenges arise in environmental science in particular because of its integrative, data intensive, interdisciplinary and collaborative nature. This thesis examines the problem of making sense of such complex and diverse sensor data in the field of environmental science, including understanding the long tail and geospatial characteristics of the environmental data.

The author postulates that Semantic Web technologies and associated techniques have the potential to address the aforementioned data challenges and support environmental science. The Semantic Web is defined by Tim Berners Lee as [9]:

"*The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*."

The vision of the Semantic Web is to shift the current World Wide Web from the medium of documents, designed for human consumption, to the medium of data and information so that computers can understand and process information without human intervention. In order to achieve this vision, the Semantic Web introduces several technologies and techniques briefly summarised below:

- Semantic annotation of data provides machine-readable and machine-interpretable metadata about different resources. The process of semantic annotation attaches additional meaningful information to different data resources. However, it does not make data machine understandable. Thus, it would require additional intelligent methods and effective reasoning and processing techniques for seamless data integration [10].
- Linked data is a mechanism that provides a set of best practices for publishing and interlinking structured data on the web [11]. It is defined as, "*data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets.*" [12]. This is a paradigm to improve an integrated mechanised access to and processing of datasets. Hence, applications can retrieve data easily across the web, irrespective of the underlying format. It can be exploited to retrieve data from multiple distributed repositories.
- An ontology is a formal specification of a shared conceptualisation [13]. It represents knowledge of a particular domain, comprised of concepts, their properties and the relationships between them. Ontologies introduce machine-interpretable meanings across different datasets. New facts and knowledge can be inferred from existing concepts/ontologies using software like reasoners that provide support for inferencing and deducing new knowledge.
- Before accessing and sharing the data first, a consistent underlying data model is required to represent data in a standard common structured format. This data model is called the Resource Description Framework (RDF). RDF represents data in the form of a statement called a triple which consists of a subject, a predicate

and an object. A collection of triples, usually stored in a triplestore, is expressed as a directed labelled graph.

- Finally, in order to access and search effectively through these triples, we need a common query language. SPARQL, a recursive acronym for the SPARQL Protocol and RDF Query Language, performs this task.

In short, with the help of Semantic Web technologies we can potentially: introduce well-explained and machine-encoded definitions of the vocabularies, integrate different datasets, deduce new facts from the existing ones and resolve the issue of data heterogeneity among the data.

The author further postulates that Semantic Web technologies have not been realised at its full potential in the field of environmental science. There have been some interesting examples, including information retrieval and management, data discovery, resolving data heterogeneity, data integration and scientific analytical workflows [14-19]. However, compared to other areas of science, the uptake of these technologies in environmental science is lower. In addition, environmental science brings major opportunities but also unique challenges to data scientists because of its inherent nature of complexity, data diversity, interdisciplinarity, and scale. Hence, there is a need for further research into the characteristics of environmental science and also how to adopt or adapt Semantic Web technologies and associated techniques in this area.

This work is carried out in the context of the Environmental Internet of Things project [20], an EPSRC-funded collaboration between Lancaster University, the Centre for Ecology and Hydrology (CEH), the University of Bangor and the British Geological Survey (BGS). The goal of the project is to design, deploy and use an IoT infrastructure for environmental monitoring and management in real-life conditions. The IoT infrastructure, deployed ‘in the wild’, examines a range of environmental facets in a particular catchment in North Wales, around the Conwy valley. The author focusses on designing and developing both a semantic data model for this project and a set of Semantic Web techniques which could represent environmental data in a potentially more unified, sharable, intelligent and reusable way. Hence, environmental science can be a good test bed for Semantic Web technologies.

## 1.1 Research Goals and Questions

The main goal of this research is to examine the potential role of Semantic Web technologies and their applicability in supporting a deeper understanding of the natural environment as derived from a plethora of sources of environmental data. This goal can be further divided into the following more specific objectives.

- Exploring particular characteristics of environmental science from the perspective of the underlying data stemming from the long tail of environmental science
- Designing a semantic data model to represent environmental data stemming from the Environmental IoT [20] data, including capturing the complex interrelationships across disparate datasets representing different environmental facets and their impact on each other
- Exploring the role of Semantic Web technologies and associated techniques to achieve such a semantic data model offering semantically enriched sensor data for performing interoperability, data integration and spatio-temporal reasoning over geospatial data, and identifying strengths and limitations of this approach
- Evaluating the overall approach through real-world scenarios/use-cases derived from the analysis of the literature coupled with semi-structured interviews carried out with leading environmental scientists

The overarching research questions that drive the research are then:

- What are the particular characteristics of data associated with environmental science, and what are the associated data challenges in terms of making sense of that data?
- What is the role of Semantic Web technologies in building a data model for the Environmental IoT Infrastructure to represent its data in all its complexity?
- What implications does this have for a technological infrastructure underpinning environmental science to exploit the potential of streaming data from IoT technology?

## 1.2 Research Methodology

In this work, the methodology would proceed along the following three phases (Figure 1.1), representing a mixed methods approach blending semi-structured interviews and experimental development.

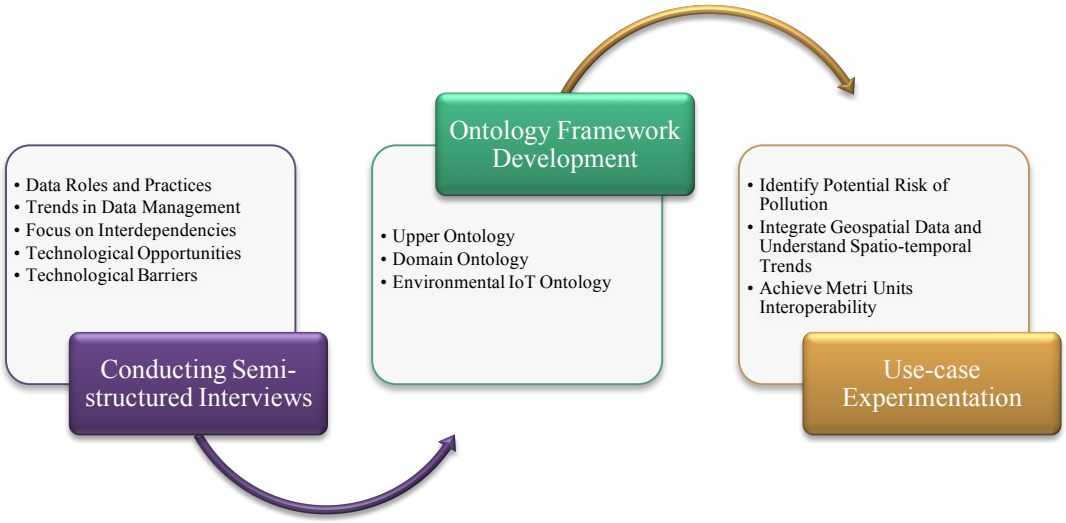


Figure 1.1: Phases of Methodology

### 1.2.1 Phase I: Conducting In-depth Semi-structured Interviews with Domain Experts

To gain an insight and knowledge of the unique characteristics of environmental science and to explore deeper the data challenges faced by environmental scientists, a series of in-depth semi-structured interviews will be conducted with domain experts. The domain experts will be chosen due to their experience and considerable expertise in their discipline. Additionally, they will be at the forefront of data-driven environmental research. Semi-structured interviews will be used for the following reasons. Firstly, this approach supports predetermined but open-ended questions in order to allow a fair degree of freedom and flexibility, allowing new questions to emerge from the dialogues. Secondly, semi-structured interviews allow the interviewer to delve deeply into the topics so that detailed knowledge of the domain is gained. Finally, this technique keeps the interview focused, conversational and allowing two-way communication. The interviews are planned to contain a number of questions covering five categories i.e. data role and practices, trends in data

management including openness, collaboration, and integration, focus on interdependency, technological opportunities and technological barriers. Some of the key findings are then fed into the later phases of the research, around use-cases (Figure 1.2).

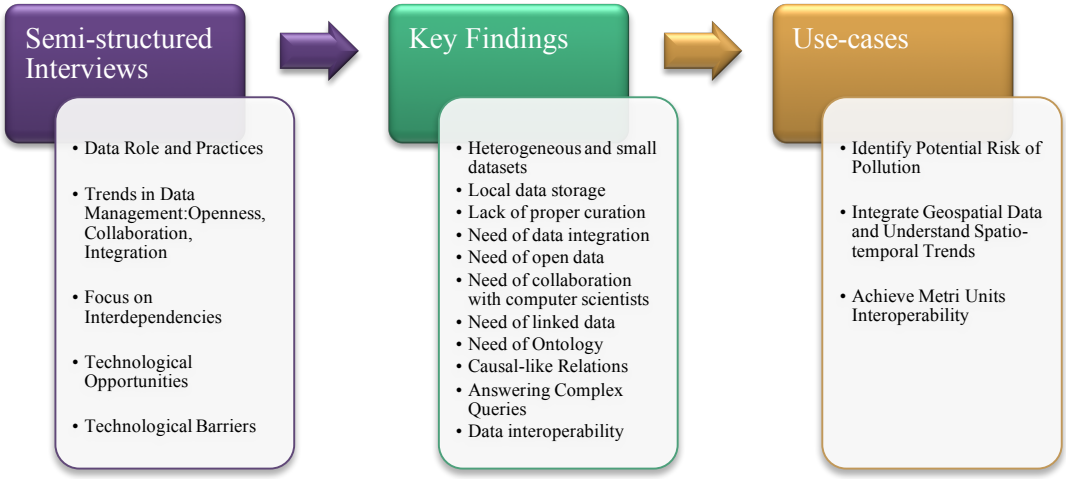


Figure 1.2: Some of the Key Findings from the Semi-structured In-depth Interviews

### 1.2.2 Phase II: Ontology Framework Development for the Environmental IoT Data

The goal of the ontology framework development is to represent different concepts and characteristics of the target domain, the relationships between them and then transforming environmental IoT data accordingly. Thus, the real-time data has to be semantically enriched with the vocabulary used in the ontology. A collaborative and incremental approach is proposed to build an ontology for the target domain of the natural environment. It is collaborative because, during the ontology design process, the input of environmental scientists will be required. It is incremental because an initial version of ontology will be developed from the domain knowledge that would have been acquired in the previous phase. The ontology will be evaluated with real-time use-cases. To conceptualise the related characteristics (such as temporal, spatial, and thematic) of environmental data, not covered by the initial ontology, it will be further modified by adding new concepts and evaluated. This process will repeat until an improved ontology is achieved. The proposed ontology framework in this work will adopt the generic model introduced by Guarino [21], which provides a top-down

approach for developing ontologies according to the level of ontological generality. Guarino's model is based on modular design that provides an easy integration of different ontologies making it suitable to be adopted in this work. The target ontology is an integrated model, which will be comprised of an upper ontology, a domain ontology and an application ontology, collectively called Environmental IoT Ontology, as shown in Figure 1.3.

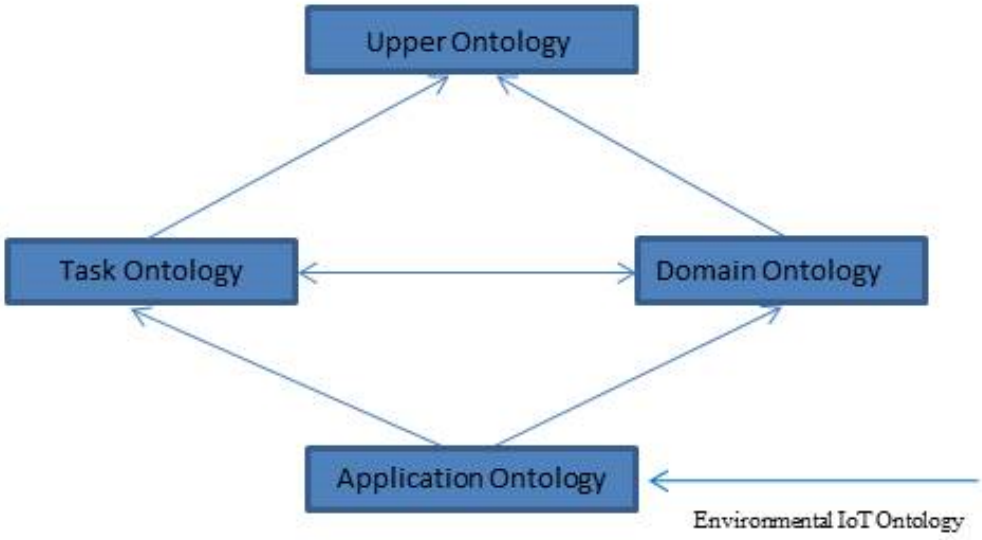


Figure 1.3: Environmental IoT Ontology Framework

### 1.2.3 Phase III: Use-cases Experimentation

After the qualitative analysis of the semi-structured interviews and drawing upon the main key findings, three key use-cases will be developed (Figure 1.1). These use-cases will be based on real-time data captured by the Environmental IoT Infrastructure [20]. They will be evaluated to test and enhance the applicability of both ontological and application framework. An iterative approach will be again adopted to incrementally enhance the ontology. This phase is iterative because, first the environmental data will be semantically enriched by the ontology, developed in the previous phase. This process of data enrichment by ontology is called data transformation. The data along with the ontology will be fed into the application framework to evaluate the real-world use-cases. If the desired results are not achieved, the process will go back to the ontology development phase so that it is modified. It will be followed by changes in the transformation of data to reflect the modified

ontology. Both the transformed data and the modified ontology will be again fed back into the application. Thus, the ontology would improve after every iteration and the process will repeat until the results are achieved.

## 1.3 Research Contribution

The thesis leads to the following contributions.

### 1.3.1 Characteristics of Environmental Data

The thesis provides some key insights into the nature of environmental data related to the long tail of science and the particular challenges associated with this area of science. These challenges include: i) discovering interdependencies between disparate datasets representing different environmental facets; ii) geospatial data integration and reasoning; iii) data heterogeneity; iv) data discovery and access; v) data quality and provenance.

### 1.3.2 Current Practices in Environmental Science

The thesis also contributes insights into current practices in data management in environmental science, including an important exploration of technological opportunities and barriers. Perhaps the most important result from this study though is the need for cross-disciplinary dialogue between environmental science and computer science so that technological opportunities can be delivered and barriers overcome.

### 1.3.3 Role of Semantic Web Technologies in Environmental Science

Through the iterative development of an ontology for streaming environmental data, it shows that Semantic Web technologies have a significant role to play in overcoming three key challenges including:

- Interdependencies between disparate datasets, overcome by semantically enriching those low-level sensor measurements using the ontology and then reasoning over the resultant enriched datasets deriving new knowledge
- Geospatial data integration and reasoning issue, resolved by again semantically enriching all sensor measurements using the ontology

- Interoperable metric units conversion, addressed by semantically assigning all sensor measurements their associated metric units using the ontology and then performing translation through inference rules.

The overall ontology is also a contribution in its own right providing a proof of concept of how a given ontology can address the needs for a given environmental project, in this case dealing with streaming data from an Environmental Internet of Things [20] deployed in North Wales.

### 1.3.4 Implications for Technological Infrastructure

The experimental work in this thesis provides extra insights into the technological needs of environmental science and in particular the underlying infrastructure needed to support scientific discovery. In particular, this thesis shows how existing technologies including ontologies, RDF, OWL, linked data and SPARQL are successfully used in underpinning environmental science around IoT data.

## 1.4 Thesis Outline

**Chapter 2** provides a background overview of Semantic Web technologies and explores the state-of-the-art on the use of such technologies and techniques in the context of eScience. The chapter provides a more in-depth assessment of related work and concludes with the argument that there is pressing need to apply Semantic Web technologies for IoT/streaming data in the natural environment because there is limited research at the intersection of the said three areas and hence further research is required particularly in terms of meeting the needs of environmental science.

**Chapter 3** examines the unique characteristics of environmental science in the context of environmental data, through semi-structured in-depth interviews. The chapter aims particularly at exploring and collecting qualitative data covering different aspects including: the role of data and practices, data trends, interdependence between disparate but interlinked datasets, and technological opportunities and barriers in environmental science. The chapter provides the analysis of the qualitative data using the Ground Theory methodology and concludes with the key findings, some of which are fed into the later phases of the work, around use-cases.

**Chapter 4** introduces the ontological framework for the environmental IoT data. The chapter provides an overall design of the ontology as well as the integration of other ontologies imported and extended in this work. The chapter concludes with the argument that the ontology in environmental science should aim for more lightweight but extensible model that communities can agree with and which can be extended over time as concepts are deemed missing.

**Chapter 5** provides an evaluation of the work through three different real-world use-cases, derived from the analysis of the semi-structured interviews. The evaluation is carried out to demonstrate the applicability and limitations of these techniques in the target discipline(s) of environmental science.

**Chapter 6** presents concluding remarks, highlighting the major contributions of the research and discussing future work. In addition, the chapter reviews the research goals and questions that have been addressed in the thesis.

# 2 Background and Related Work

## 2.1 Introduction

The World Wide Web has been evolved from the Web of documents to the Web of data (the Semantic Web) with the vision to create a globally connected data space [22]. The Semantic Web has been applied in various fields where there is a wide deployment of heterogeneous information of different quality, for instance eScience [23]. The need for Semantic Web technologies in environmental sciences has been growing and has already gained acceptance in other fields such as solar-terrestrial physics [24-25], ocean and marine sciences [26] and health care and life sciences [27-28]. Because of the growing need of shared semantics and the heterogeneous nature of environmental data, environmental science can be a good test bed for Semantic Web technologies.

The main goal of this chapter is twofold: to review technological developments and to assess the state-of-the-art in Semantic Web for environmental science. To place this work in context, the chapter also offers a broader perspective on science, introducing eScience and its related trends including open science and the fourth paradigm of science.

This chapter is structured as follows. Section 2.2 provides a background on eScience/cyberinfrastructure and its related trends. Sections 2.3 provides an overview

of the underlying Semantic Web technologies. Section 2.4 provides a more in-depth analysis of the related work. Finally, Section 2.5 provides an analysis of the state-of-the-art and concludes with the argument that there is pressing need to apply Semantic Web technologies for IoT/streaming data in the natural environment because there is limited research at the intersection of these three areas and hence further research is required particularly in terms of meeting the needs of environmental science.

## 2.2 Background on eScience

### 2.2.1 Introduction

The Internet has played an overarching role in the advancement of modern science which has become more complex, rapidly scalable and increasingly dependent on data [29]. Because of this large scale, complex and data intensive nature of science, it demands more distributed, collaborative and interdisciplinary research groups [30] so that scientists could process and share their data, experiments and results. To undertake scientific research in this new paradigm, computer scientists need to develop advanced scientific, methodological, and computational information processing techniques and a new powerful supporting cyberinfrastructure over the Internet [31]. To refer to such computing infrastructure, a new term 'eScience' was introduced in the UK to enable scientific exploration accomplished through world-wide collaboration and multidisciplinary and interdisciplinary research (with an equivalent term 'e-Infrastructure' used in Europe and Cyberinfrastructure in the US) [31].

The idea of doing collaborative research on the Internet can be traced back to William Wulf's vision of 'collaboratory' in 1989 [32]. He coined this new term by combining the words collaboration and laboratory and defined it as a:

"*Centre without walls, in which the nation's researchers can perform their research without regard to geographical location- interacting with colleagues, accessing instrumentation, sharing data and computational resource, and accessing information in digital libraries*."

The term eScience was first coined in 1999 by Dr. John Taylor, then Director General of Research Council in the UK Office of Science and Technology (OST) [33]. He defined the term as:

"*eScience is about global collaboration in key areas of science and the next generation of infrastructure that will enable it*."

He also claimed:

"*eScience will change the dynamics of the way science is undertaken*."

The term cyberinfrastructure was first used in the NSF's 2003 final report, also called the 'Atkins Report' entitled "Revolutionising Science and Engineering through Cyberinfrastructure" [34]. The report defines infrastructure vis-à-vis cyberinfrastructure as:

**"***The term infrastructure has been used since the 1920s to refer collectively to the roads, power grids, telephone systems, bridges, rail lines, and similar public works that are required for an industrial economy to function. Although good infrastructure is often taken for granted and noticed only when it stops functioning, it is among the most complex and expensive thing that society creates. The newer term cyberinfrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy*".

The NSF's Cyberinfrastructure Council 2007 report, titled, 'Cyberinfrastructure vision for 21st century discovery' [35], defined cyberinfrastructure as:

"*Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information and knowledge*".

The types of services and facilities provided by a cyberinfrastructure layer (shaded) to enable new knowledge environments for research are illustrated in Figure 2.1 [34].

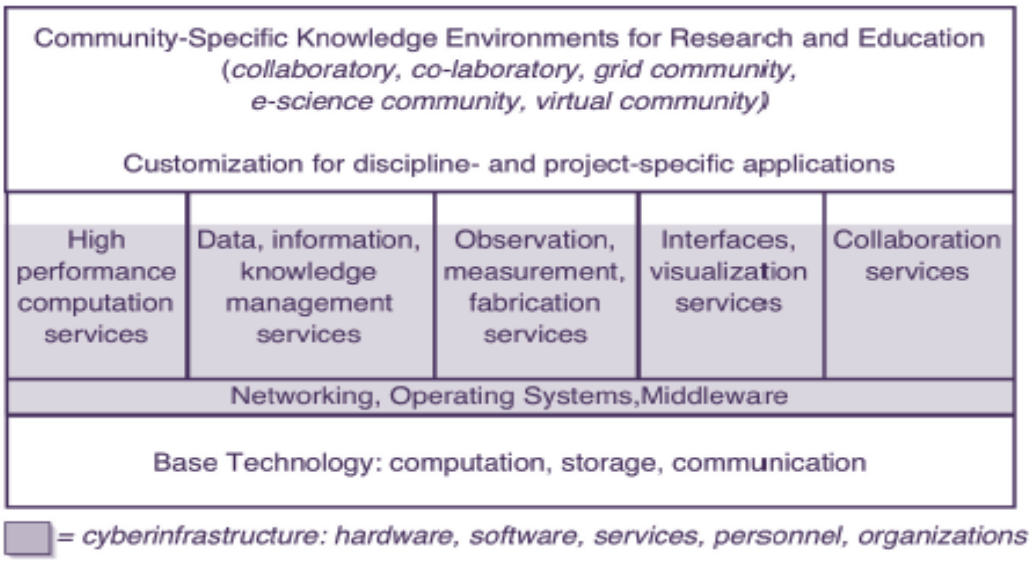


Figure 2.1: Integrated Cyberinfrastructure Services [34]

The commonalities across these views are significant, with the main focus being on salient characteristics of eScience including interdisciplinary collaboration, the data-centric nature of the science and openness [36]. Furthermore, interoperability is crucial to enable research in an interdisciplinary and open environment, where a huge amount of complex and heterogeneous data is generated.

### 2.2.2 eScience Challenges

The vision of eScience promises new prospects of undertaking scientific research through collaborative and interdisciplinary scientific processes over the Internet. Through this paradigm shift in scientific research, scientists would be able to generate, process, analyse, share and discuss their data, understanding, experiments and results in a more effective way [23]. However, to achieve this vision, some technical challenges need to be overcome. There are many challenges [37] but in the context of this thesis, the most relevant ones are summarised below:

- To meet the requirements of open data-rich information system that demands both semantic information and services to perform data processing and reasoning.

- To resolve interoperability among geographically distributed heterogeneous resources in order to fulfil the requirements of a composite scientific process.
- To attain high-quality and domain-specific metadata for automatic data integration and interpretation that plays a key role in knowledge discovery over a huge amount of data.
- To develop intelligent software applications that must be able in understanding and interpreting the correctness and right context of data and associated metadata.

There has been a serious effort to address the aforementioned challenges in order to make the e-Science vision viable. The driving force comes from the recent advancement in information and communication technology and the new computing paradigms including High Performance Computing, Grid and cloud computing. These eScience enabling technologies provide opportunities to undertake eScience research in a distributed, collaborative and integrative manner. On the other hand, to process, integrate, and analyse this huge amount of data leads to challenges including data discovery, heterogeneity, integration, to name but a few. Hence, not only is there a need for Semantic Web technologies in eScience research to potentially address the aforementioned challenges but also there needs to be the community pull supporting interdisciplinary data-driven and open research to turn the data into knowledge.

### 2.2.3 Trends in eScience

#### Open Science

Modern science is characterised by its public character which promises cooperation in research and free access to knowledge among the researchers [38]**.** According to John Ziman, scientific knowledge does not exist *"by the moral authority or literary skilsl of its creator, but by its recognition and appropriation by the whole scientific community."* [39] It aims at developing a consensus of views on the basis of facts and theories. The consensus, achieved through peer review, empirical evidences and critical analysis of highly intellectual researchers, establishes "scientific objectivity." This has led to the establishment of open science that makes the scientific information and research results open and free to the community. Open science as defined by [40] *"is the optimal sharing of knowledge and supporting tools, such as publications, research data, software, educational resources and infrastructures, across*

*institutional, disciplinary and national boundaries"*. Openness strengthens the scientific method and knowledge can be improved, or rejected through scrutiny and critical analysis [41]. Releasing scientific theories along with their experimental data to public allows them to be strictly and thoroughly examined and corrected for the errors if possible, making them refined or rejected [42]. Thus, the scientific knowledge progresses further through this open scrutiny and challenge. Open access to scientific knowledge has been practiced by many preprint servers, scientific journals, researchers' websites and worldwide institutional repositories and facilitated by Science Commons for licencing.

**The Fourth Paradigm of Science**

The data intensive science, also called "the fourth paradigm", was proposed by the Turing award winner, the late Jim Gray in 2007 working for Microsoft. Gray's vision of highly sophisticated algorithms and tools to visualise, mine, analyse and manipulate scientific data can bring solution to the complex research problems of modern science [43]. The first two paradigms of scientific discovery, experimentation and theory which have been dominant for centuries have a long history. Experimental science goes back to ancient Greece and China, when people used observations, descriptions and experimentations to do science. The second paradigm is that of developing a theory to explain a new phenomenon of natural world such as Newton's theory of gravitation and laws of motion and Maxwell's equations etc. With the advent of modern high performance digital computers in the latter half of the 20$^{th}$ century, the third paradigm of science, computation and simulation for scientific discoveries, was introduced by the Nobel Prize winner Ken Wilson. These extensive simulations enabled the scientists to discover those areas of discovery which were difficult to reach by experimentation and theory such as weather forecasting, climate modelling and galaxy formation. The fourth paradigm of science, also called 'Big Data Science' does not replace the other three methodologies but demands for a distinct set of skills. This paradigm exploits the large volumes of data generated by simulations or sensor networks and processed by advanced software tools for visualisation, data mining and statistical analysis to progress the scientific discovery process (as shown in Figure 2.2) [44].

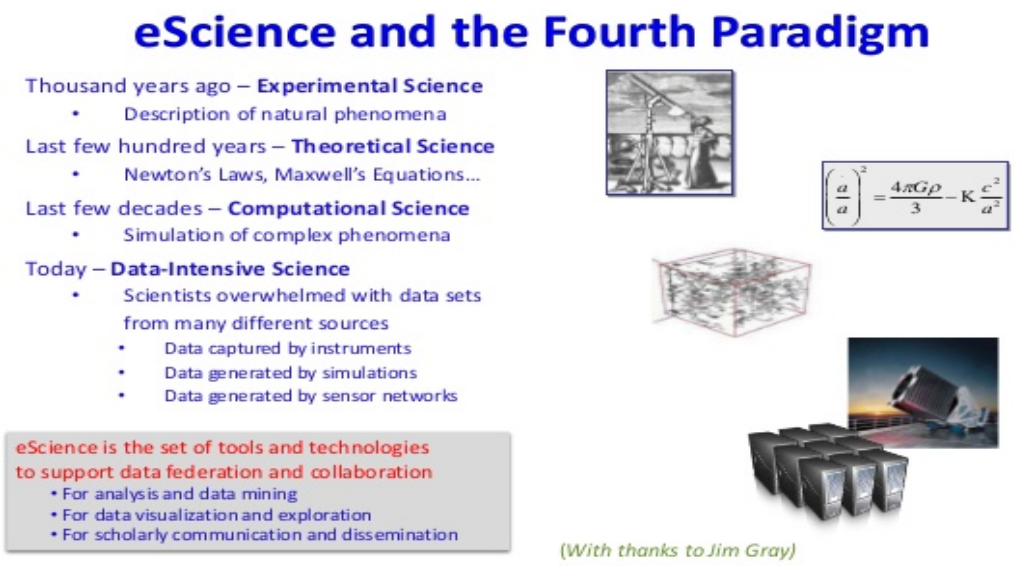


Figure 2.2: Paradigms of Science [44]

## 2.3 Background on the Semantic Web

### 2.3.1 Introduction

Tim Berners-Lee introduced the idea of the Semantic Web in his keynote at the first World Wide Web conference in 1994 [45]. A few years later, he expressed the vision of the Semantic Web as:

"*I have a dream for the Web [in which computers] become capable of analysing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialise.*"

This vision was developed further in his first article published in Scientific American in May 2001 [9]. In the aforementioned article, he defined the Semantic Web as:

"*The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*"

The vision of Semantic Web is to shift the current World Wide Web from the medium of documents, designed for human consumption, to the medium of data so that computers can understand and process information without human intervention. The reasons why this shift is required are the facilitation of reusing the data in new context, the alleviation of costly information extraction from documents done by humans, and the release of vast amount of relational database tables and spreadsheets data, presently inaccessible, through automatic processing by machines [46].

### 2.3.2 Underlying Technologies

This section provides an introduction of Semantic Web technologies.

**(a) Resource Description Framework (RDF)**

One of the main problems of the World Wide Web is that it only supports human interaction; in other words, it is primarily built for human browsing and searching HTML documents [47]. This model is lacking in precision and is inadequate for browsing a huge amount of information to locate the desired document rapidly because it searches the documents on the basis of text string matching. Thus, the current model of web search and information retrieval is inefficient in looking for the required web documents. Furthermore, the information extraction from documents by humans involves mental fatigue. Therefore, it has been proposed that we need a framework based on metadata which enables the description of web documents in a more precise manner, to enhance the web search efficiency and precision and turn the current web of documents from machine-readable to machine-understandable [48]. More specifically, the Resource Description Framework (RDF) model has been proposed to provide the necessary underlying support for the above challenges. In addition, it provides interoperability among web applications that transfer machine-understandable information.

RDF [49] is a data model and XML-based language that represents information in the web and enables data integration by resolving semantic differences. It is a metadata framework and a knowledge representation scheme that provides encoding, exchange and reuse of structured metadata [50]. Through RDF, we can publish both human-readable and machine-processable vocabularies which are developed in order to support the reusability and extension of metadata semantics among different

information groups. It also allows metadata interoperability among different metadata frameworks. It provides a syntax independent representation to describe web resources. A resource is an object which can be anything in the world such as a web page, a web site, or anything having some information about something. Every resource is recognised by a unique identifier called Uniform Resource Identifier (URI). Resources have attributes which are described by property names and their corresponding values. Values might be either atomic (text, strings, numbers, et.) or other resources having their own properties. A collection of properties describing the same resource is called a description. Thus, RDF has three main components i.e. resources, properties which describe a resource and a statement which is a combination of a resource, its properties and their corresponding values. These three individual components of a statement are also known as subject, predicate and object respectively. These RDF triples can be expressed through a graph notation with nodes representing web resources and labelled edges representing properties. RDF has a number of application areas such as resource discovery, content cataloguing, electronic commerce, intelligent software agents, digital signatures, content rating, intellectual property rights and privacy preferences and policies etc.

**(b) Ontology**

The concept of an ontology was coined in 1613 and its origin dates back to Aristotle. In philosophy, it is defined as *"the study of being"* or *"the study of what might exist"* or *"the subject of existence"*. In other words, it is a branch of philosophy that deals with the nature of existence. In the context of computer science, Thomas Gruber defined an ontology as [13]:

"*In the context of knowledge sharing, I use the term ontology to mean a specification of a conceptualisation. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy*."

Gruber described the idea of conceptualisation in accordance with Genesereth and Nilsson [51] who said: "*A body of formally represented knowledge is based on a*

*conceptualisation: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. A conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledgebase, knowledge-based system, or knowledge-level agent is committed to some conceptualisation, explicitly or implicitly*."

In 1997, Borst, with a little modification to Gruber's definition, defined ontologies as: *"Ontologies are defined as a formal specification of a shared conceptualisation."* [52]. In 1998, Studer et al. [53] combined these two (Gruber and Borst) definitions and defined ontologies as: *"An ontology is a formal, explicit specification of a shared conceptualisation. A 'conceptualisation' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine readable, which excludes natural language. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group."*

Some researchers take Gruber's article as the beginning of ontology research in computer science but its role in Artificial Intelligence for knowledge engineering goes back to the 1980's article by John McCarthy [54] followed by Hayes [55] in 1985 and Alexander et al. [56] in 1986. Alexander et al. for the first time, presented a knowledge engineering methodology, called ontological analysis. They developed a family of languages collectively called SPOONS (SPecification of Ontological Structure) that encompassed tools based on domain equations, equational logic, and semantic grammars respectively. This was perhaps the first departure of ontology from philosophy to computer science; that is taking it from the nature of existence to the collection of abstract objects, relationships and transformations in order to use it as an AI tool for knowledge engineering in a particular domain of interest. Since then, the ontologies have been played a key role in information systems, natural language understanding, knowledge based systems, database design, software engineering and the Semantic Web.

**Why do we need to develop ontologies?**

One of the main purposes behind ontology development is that it plays an important role in information sharing [57] among people or software agents. For instance, in the medicine field, the Unified Medical Language System is a large, standardised structured vocabulary which can be used by software agents to share, extract and aggregate medical information with other applications or answer user queries.

Noy et al. [58] described other important reasons which are briefly described.

- Ontologies allow reusing domain knowledge which makes it one of the primary reasons in rushing into ontology research. For example, in our research, we are going to integrate and extend several existing ontologies including SSN (Semantic Sensor Network), Time, Geo, GeoSPARQL and MUO/UCUM. Similarly, in Earth Sciences, SWEET (Semantic Web for Earth and Environmental Terminology) [59] is a collection of ontologies in earth and environmental sciences that has been (re)used by other research groups doing ontology development in the same or relevant areas.
- Furthermore, ontologies help in making the domain assumptions clear and easy to understand and can be changed very easily if the domain knowledge changes.
- Ontologies enable us to separate the domain knowledge from the operational knowledge.
- Finally, ontologies enable us to analyse domain knowledge and help in clarifying the structure of knowledge which is very important in case of reuse and extending the existing ontologies [25].

**What are the different types of ontologies?**

There are different kinds of ontologies including:

- Generic or upper ontologies - capture knowledge that can be used in multiple domains. Typically, generic ontologies describe concepts including space, time, matter, state, object, event etc. [60]
- Domain ontologies which are developed for representing knowledge in a particular area of interest or domain (for example earth sciences, bioinformatics, e-commerce etc.).

- Method or task ontologies describe how domain knowledge can be used to perform specific tasks (e.g. diagnosis or selling). Methods are used to describe the functionality of an application thus application ontologies can hardly be used for other applications.
- Application ontologies are those which can be used to design an application and contain both domain ontologies and methods from method ontologies [61].

**How to represent knowledge in an ontology?**

To represent knowledge in an ontology is a design decision that requires an objective criterion in order to guide and evaluate such design. Gruber [62] suggested five ontology design principles for the purpose of knowledge sharing and interoperation among applications which are briefly described here.

- The first design criterion is the clarity of the definitions which says the meaning of the defined terms should be effective, objective, with no or less ambiguity and independent of social or computational context. All definitions should be recorded in natural language and if possible, complete definitions should be preferred over partial definitions.
- The second design principle is the coherence which says ontologies should allow only those inferences which are consistent with the definitions. Coherence should also be applicable to the informal definitions used in natural language documentation.
- The third design rule says ontologies should be extendible in order to accommodate the anticipated tasks so that one can easily extend and specialise the existing shared vocabulary without revising the existing definitions.
- The fourth principle is about minimal encoding bias which states the conceptualisation should be specified at the knowledge level irrespective of the convenience of notation or implementation.
- Finally, ontologies should need the minimal ontological commitment enough to support the desired knowledge sharing activities. Ontologies should commit as few claims as possible in order to permit other parties to specialise and instantiate the ontologies according to their needs.

**What are the application areas of ontology in Computer Science?**

Ontologies play a vital role in computer science including modelling complex areas of knowledge, resolving interoperability issues, searching large datasets and systems engineering. These application areas are briefly described below.

Firstly, in some knowledge domains, representation of knowledge is not a difficult task to describe the fundamental characteristics of well-defined and local areas of interest. Nevertheless, there used to exist some complex areas of knowledge in which knowledge representation was such a challenging task. One such example is the description of mutant phenotype which was not easy to describe it in a simple way [63]. It is defined as "the observable and measurable characteristics of an organism, which result from the interaction of the organism's generic 'blueprint' (its genotype) and the environment." In most biological databases, phenotype information was stored in free-text form [64-66], though some structured ways of storing information also existed, which was not easy to query and compare these free-text descriptions. This issue of phenotypic descriptions was tackled effectively through developing ontologies in different ways such as designing dedicated ontology specific for an organism, or through a composite annotation using several simpler ontologies, or by combining the defined terms in multiple orthogonal ontologies to create a single new ontology.

Secondly, another promising application of ontologies in computer science and information science is the provision of interoperability support gained by translating between different modelling methods, computing paradigms, languages, representations and software tools. The researchers in Semantic Web community usually tackle the problem of interoperability on the basis of reasoning principles or inference rules, using ontologies as a cross-cutting technology [67]. In ontologies, the knowledge base might contain effective and complete operational defined terms and the relationship between those terms; thus, one term can be expressed accurately in terms of another using equality based axioms or mappings and therefore can support more "intelligent" interoperability [68].

Thirdly, the current web is a huge semi-structured database consisting of billions of documents. It has been continuously growing rapidly over the past many years

making both information retrieval and knowledge management challenging tasks [69]. With the deployment of ontologies in Semantic Web applications, information retrieval has become very effective to a great extent. Ontology is a best means of arranging or organising an information repository and can be used as a sophisticated indexing mechanism in order to facilitate searching large datasets [70]. Information repositories, structured on the basis of ontologies and semantic annotations, add meaning to the web pages, thus refine and aid web search. The inference engine, using background ontologies, further enhances these semantic annotations on the basis of inference rules. Hence, it adds all properties that can be deduced/induced from the semantic annotations and ontologies [71].

Finally, ontologies have also drawn attention from software engineering community, where the software engineers design the ontology to characterise and specify the entities of a knowledge domain and use it as a base for software specification and development [72]. For example, ontology can be used as a reusable or shared component in an application to achieve software reusability; it can perform consistency checking on the basis of properties and value restrictions to develop more reliable software; it can help in guiding knowledge acquisition and designing the software requirements and specification document for a knowledge-based systems; moreover, ontology-based systems also help in improving software documentations which result in reduced software maintenance cost.

**(c) Linked Data**

Linked data is a mechanism to describe a set of best practices for publishing and interlinking structured data on the Web. It is defined as [11]:

"*To make the Web of Data (Semantic Web) a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but relationships among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets). This collection of interrelated datasets on the Web can also be referred to as Linked Data.*"

Bizer et al. [12] defined linked data as:

"*Linked data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets.*"

Berners Lee described the significance of linked data as [22]:

"*The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.*"

**What is the rationale of Linked Data?**

Linked data plays a key role in sharing and reusing data on the Web. The main factor in data reusability is to what extent it is structured [73]. If the structure of data is well defined and regular, it can easily be processed by different application tools for reuse. As the Web documents in the classical Web are unstructured or loosely structured, software applications find it very difficult to extract meaning from HTML pages and could use it for smart purposes. One of the solutions to resolve this issue is microformats [73-74] which promote publishing structured data on the Web by embedding data about people, organisations, events, reviews and ratings in HTML pages through class attributes. The downside of microformats is the support of limited number of different types of entities, attributes describing these entities, and often the inability of expressing relationships between entities because of having no identifiers. The second mechanism to provide structured data on the Web is through Web APIs which enable access to data through querying over the HTTP protocol [73-75]. A couple of well-known examples of Web APIs are the Amazon Product Advertising API (http://docs.amazonwebservices.com/AWSECommerceService/latest/DG/) and the Flickr API (http://www.flickr.com/services/api/). Thousands of Web APIs are maintained in a directory by a website named ProgrammableWeb [76]. Web APIs resulted in numerous specialised web applications such as mashups that combine contents into an integrated experience from more than one source; each of which is accessed through a public interface or API. Though Web APIs provide a number of advantages to access structured data on the Web, still this mechanism has some serious shortcomings [75]. First, these APIs provide proprietary interfaces and cannot be accessed using generic data browsers. Second, they fragment the Web into separate

data silos and mashup developers are restricted to fixed set of data sources. Finally, the scope of Web APIs' identifiers to refer to data items is local, hyperlinks can't be set between data objects provided by different APIs. Consequently, the data in the Web is not linkable and discoverable at their full potential.

To overcome these problems, Tim Berners Lee introduced four main rules in his Web architecture note entitled 'Linked Data' to publish and interlink structured data on the Web [22]. These practices are also known as Linked Data Principles which are described as under:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

**What are the advantages of Linked Data?**

Linked data provides some promising benefits discussed below [77].

- Linked data relies on RDF which is particularly designed for global data sharing. In RDF, information is expressed by unique identifiers called URIs. Hence, linked data provides a unifying data model.
- By using RDF, it enables syntactic and semantic data integration of different linked datasets through schema and instance matching techniques and by relying on shared vocabularies and ontologies and connecting different definitions through vocabulary links.
- It provides coherence in which data items, represented by URIs in a triple (from different namespace) are effectively interlinked.
- It provides a standardised data access mechanism by using a world-wide standard HTTP protocol, thus allowing generic data browsers for accessing data and search engines for crawling the global data space.
- It provides data discovery at runtime by using URIs to connect different data sources and following RDF links to create a global data graph.

**Where does Linked Data apply?**

There are numerous applications that leverage the web of data which can be categorised into linked data browsers, linked data search engines and specialised applications.

Linked data browsers enable users to surf the web of data by following links in RDF triples. New data can be discovered and merged automatically through owl:sameAs links. Examples of linked data browsers are Tabulator [78], Marble [79], Disco-Hypermedia Browser [80], Fenfire [81], and Humboldt [82] etc.

Linked data search engines that crawl Linked Data through RDF links are of two types. One, which is human-oriented, serves users on keyword basis and follows the interaction mechanism of Google and Yahoo, includes Falcons [83], SWSE (Semantic Web Search Engine) [84]. Another category is application-oriented Indexes which serve the requirements of other applications through APIs, includes Swoogle [85], Sindice [86] and Watson [87].

Linked Data specialised applications that are developed to serve particular domain include DBpedia Mobile [88], a location-aware Linked Data browser developed for smart phone users to discover a city; Revyu [89], a reviewing and rating website to help users improve their experience; and Talis Aspire [90], a web-based resource list management application developed to help university lecturers and students.

**What is the Linked Data lifecycle?**

Soren Auer et al. [91] describe different stages involved in the linked data lifecycle as illustrated in Figure 2.3. The steps involved in the lifecycle need not be sequential. These stages are summarised below.

**Extraction**- The first step in Linked data lifecycle is the information extraction in which the information is mapped from unstructured (e.g. text), semi-structured (e.g. XML), and structured (e.g. relational tables) representations to the RDF data model.

**Storage/Querying**- Once sufficient RDF triples are gathered, the next step is to store these triples and query them efficiently through a querying language.


Figure 2.3: The Linked Data Lifecycle

**Authoring-** Here the users create, modify, and extend the structured information by exploiting some Semantic Wiki technologies such as OntoWiki.

**Linking-** Perhaps the most important concept in the Semantic Web is linking between entities if the information provided by different data publishers refers to the same or related web resources.

**Classification/Enrichment-** to transform linked data from raw form into a regular structure, schema and classification for efficient data integration, querying and search purposes. Through enrichment methods (e.g. reasoning), we can increase the expressiveness and semantic richness of a knowledge base.

**Quality Analysis-** mechanisms to assess the quality of data (if it is inconsistent, incomplete, inaccurate or obsolete) on the basis of different parameters such as provenance, context, and structure etc.

**Evolution/Repair** - ensuring transparency when changes occur to knowledge bases, vocabularies and ontologies and fixing them if problem arises in result of those changes.

**Search, browsing and exploration** - developing better techniques for searching, browsing, exploring and visualisation to use linked data efficiently and easily.

### 2.3.3 Summary

Semantic Web technologies are emerging in underpinning environmental science to understand this multi-disciplinary, integrative and data-driven science. Various eScience areas (most notably disciplines include health care and life science) are much further on accepting Semantic Web technologies. Furthermore, the Semantic Web community has widely focused on formal aspects of semantic representation languages or general-purpose semantic application development. However, as mentioned in the previous chapter, they have done little research to address the data challenges in the natural environment. This little uptake leaves a semantic gap in (a) understanding highly complex and heterogeneous environmental data (b) turning this underlying data into knowledge and (c) integrating and interlinking it with other data sources to make a unified view of the data (and by exploration knowledge). Hence, there is a need to further explore these technologies to understand the characteristics of this integrated and data-driven science around data in all its complexity. The next section therefore looks in more detail at the related work in these technologies to determine the current state-of-the-art.

## 2.4 Related Work

### 2.4.1 Dimensions of the state-of-the-art

To perform a systematic comparison of related initiatives and developments, this section introduces a set of dimensions in order to capture key features in a consistent manner. These dimensions are described below.

i) The purpose of the ontology: The main purpose of the ontology is to capture knowledge of a particular domain in order to enable semantic applications and machines to better understand the target domain and the relationships among different concepts of the domain. This dimension is important in the context of the Environmental IoT project and beyond to develop an ontology for describing data and also capture complex interrelationships across disparate datasets representing different environmental facets.

ii) The coverage of the ontology: The coverage, also called scope, of the ontology determines the potential maximum range of concepts describing a particular domain. In the context of a semantic sensor network, an ontology may specify sensor descriptions, observations and measurements. An ontology can be either mainly sensor-centric or observation-centric or both. The coverage of ontology is very significant in this work because not only the sensor ontology should describe sensors and observations but other important features including thematic, spatial and temporal dimensions of the domain should also be modelled.

iii) Expressiveness of the ontology: The expressiveness dimension demonstrates the ability of an ontology language to capture certain aspects of a particular domain. More expressive ontology languages can conceptualise a large variety of knowledge about a domain, however at the cost of computational complexity. This dimension is significant in the context of this work because a sufficiently rich language is required to capture a wide variety of concepts while at the same time preserving efficient reasoning support.

iv) Using existing standards: One of the main reasons of ontology development is that others can use the existing standards to save time and efforts. Using and instantiating existing standards also help in the provision of interoperable solutions. This dimension is taken into account to both adopt and adapt existing standards to achieve portability and semantic interoperability on a wider scale.

v) Semantic annotation of data: Semantic modelling and ontologies attach additional meaningful information to data resources to provide machine-interpretable descriptions. Semantic annotations of sensor data and IoT devices using sensor and domain ontologies is necessary in this work in order to support querying, searching and reasoning over environmental data in a sensor network.

vi) Semantic data integration: In the context of a semantic sensor network, data usually stem from a variety of sources and hence requires combining it with other data sources to facilitate context awareness. This dimension is essential because it enables environmental scientists to form a unified view of the structure and more importantly semantics of heterogeneous environmental data.

vii) Semantic reasoning: The Semantic Web technologies formalise knowledge in a way that enable reasoning over data that is implicitly declared to infer new knowledge. Semantic reasoning in the context of IoT data for the natural environment is an important tool to derive high-level knowledge from low-level sensor

measurements, for instance, deducing the risk of a pollution, soil saturation or a storm event.

viii) Semantic interoperability: The exchange and interpretation of data in an unambiguous way by different software and machines to support automated or semi-automated interaction. In the context of IoT data for the natural environment, providing interoperability is one of the most important dimensions owing to the issues of heterogeneous nature of devices, data models, and software tools.

ix) Effective querying support: Once the data is semantically enriched and stored in a triplestore, users require access to and searching the data effectively to enhance further interaction with the resources. This dimension is taken into account because extended and effective querying support would be required to address the complex questions of users (scientists) in the target domain.

### 2.4.2 Survey of the state-of-the-art

This section applies the aforementioned dimensions to survey the related work in the area of the Semantic Web, particularly ontology design specifically for IoT/streaming data for the natural environment, as shown in the diagram (**Figure 2.4**, marked in red). The section surveys the related work by examining research in the different regions of the diagram with emphasis on work that lies at the intersection of the areas of: a) Semantic Web for IoT data (marked in black); b) Semantic Web in environmental science (marked in purple); and c) Semantic Web for IoT data in environmental science (marked in red). The work that lies at the intersection of IoT/streaming data in environmental science (marked in green) [92-108] is mostly technology-oriented focusing on issues related to resource-constrained IoT devices and communication, and hence is beyond the scope of this thesis.

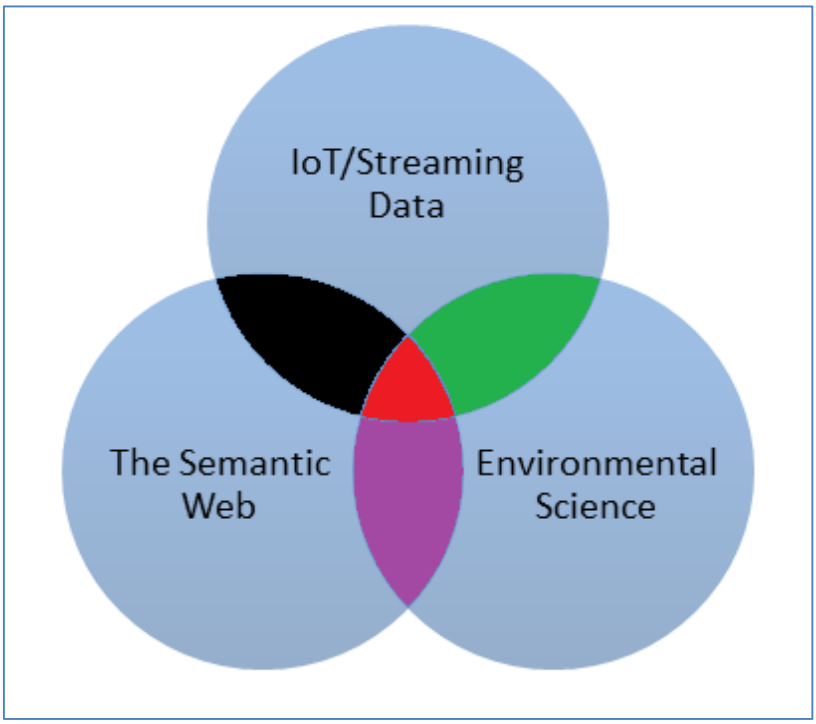


Figure 2.4: Target Area of Research (marked in red)

**(a) The Semantic Web and IoT/Streaming Data**

The Internet of Things has become a reality today connecting billions of devices and things in numerous fields including industry, health, infrastructures and the natural environment, to name but a few [109]. One of the overarching goals of IoT by connecting these devices and capturing data from them is to create situation or context awareness, and enable applications, machines and humans to better understand their surrounding environment [10]. However, to achieve this goal, it raises some technological issues at semantic level because the data collected from these devices is diverse, heterogeneous and may be spatio-temporal. These characteristics make challenging several tasks including capturing complex interrelationships, data integration and reasoning, and interoperability. Applying Semantic Web technologies to IoT devices can potentially achieve the above-mentioned goal of IoT, provided the said data challenges are addressed. This section surveys existing work that used Semantic Web technologies including ontologies and linked data for resolving the issues of capturing complex interrelationships, data integration and reasoning, and data interoperability in a sensor network.

Ontologies have been playing an important role in addressing the aforementioned challenges in a sensor network. For instance, the work of Avancha et al. [110] proposed an ontology for sensor networks to capture important features of a sensor node including both functionality and current state. The ontology focused mainly on high level descriptions of components of nodes and functional descriptions of sensors. However, it did not provide good coverage, i.e. it gave little attention to sensors, systems and measurement procedures. The OntoSensor [111-112] ontology was built to enable applications for advanced inference methods to be used over heterogeneous sensor data. It adopted the concepts and properties from SensorML, the IEEE SUMO ontology, ISO 19115 and some constructs from the Web Ontology Language. OntoSensor covered a broad range of concepts; however, it lacks a proper data description model to provide interoperability for data representation and observation. Kim et al. [113] later extended the OntoSensor ontology for web services. Their ontology comprised three main components: ServiceProperty, LocationProperty, and PhysicalProperty. However, their system did not specify the description and interpretation of sensor data in a sensor network application. Moreover, due to poor ontology modelling of concepts it was not reused or extended in other applications. The SWAMO project proposed an ontology for an intelligent agent based framework to describe physical devices, processes and tasks [114]. Unlike the ontologies proposed by Avancha et al. that focused primarily on data and measurements, the SWAMO ontology included the systems aspect e.g. survival and operating range, and deployment, in addition to sensors and measurements. Its main benefit was providing interoperability with SensorML, and Sensor Web Enablement standards. However, the overall approach lacked cohesion that is the relatedness of elements in an ontology which measures modularity. Low cohesion can lead to modularity issues. The A3ME (Agent-based Middleware approach for Mixed Mode Environments) ontology was developed to classify the discovery of sensor devices and their capabilities in heterogeneous networks having resource constrained sensor nodes [115-116]. The A3ME ontology covered a wide range of concepts; however, it was mainly designed for low-power devices and did not support complex reasoning. CSIRO developed a sensor ontology to describe and reason about sensors, observations and scientific models [117-118]. The main objective was the usage of sensor reasoning and querying approaches for enabling data integration, searching, and classification. It was relatively an expressive ontology, however faced some issues. The processes

defined in the ontology could not express their functions and hence an external reasoning mechanism was required. Furthermore, the types of inputs, outputs and what a sensor measures could not be expressed properly in OWL. Sheth et al. [119] presented the idea of a Semantic Sensor Web (SSW) framework to give enhanced meaning and descriptions to sensor observations in order to enable situation awareness. In their paper, they explained how a Semantic Sensor Web can enable interoperability, advanced analytics and reasoning over heterogeneous sensor data for situation awareness by using semantic annotation, ontologies and rule-based reasoning. However, their work focused on achieving interoperability between sensors rather than data. Besides, their writing does not explain how SSW could relate to existing knowledge on the Web. To understand and conceptualise the information processes involved in observations, Kuhn in [120] proposed a general ontology to formalise the semantics of observations. The ontology modelled both human and technical sensors and the role of an observer in order to cover a wide range of current and evolving Semantic Sensor Web standards. It represented a first step towards an ontological foundation to deal with observations; however, it did not identify reasoning requirements for sensor data integration.

Building on the experience of this work, the W3C Semantic Sensor Network Incubator group (SSN-XG) developed a general, domain independent ontology known as the Semantic Sensor Network (SSN) ontology [121-122]. The SSN ontology is based on the Stimulus-Sensor-Observation (SSO) pattern [123] and describes sensors and their capabilities, observations, systems and deployments. The SSN ontology has some important features, for instance, it is compatible with other standards including OGC SensorML at the sensor level and O & M at the observation level. Moreover, as the SSN is a generic ontology, it can be adopted in many scenarios and domains. However, the SSN ontology does not provide concepts to describe temporal, spatial, units of measurements and domain knowledge. In addition, it does not provide specifications for features or types of observed properties.

There is now a body of work on using or adopting the SSN ontology in various areas of applications. Gray et al. [124] described Semantic Sensor Web architecture to discover and integrate multiple heterogeneous datasets. The good feature about their architecture lies in the provision of support for semantic sensor web applications both

for discovering and integrating spatio-temporal and thematic data. However, the approach lacks reasoning capability over sensor data. Wang et al. in [125] developed a lightweight semantic description model for knowledge representation in the IoT domain and reused some existing ontologies including SSN. The design of their ontology followed the recognised best practices in ontology engineering and modelling. However, their approach focused mainly on service discovery, testing and dynamic composition and not on actual observation data. Barnaghi et al. in [126] used the SSN ontology in their framework for translating low-level sensor data to high-level abstractions to infer perceptions using OWL reasoner. However, their paper reported an ongoing work and the results were at an early stage. Besides, their solution was limited to predefined inference models. Roda and Musulin [127] presented an ontology-based framework, reusing SSN, to perform intelligent data integration and analysis on sensor measurements. The positive feature of their framework is the modular design enabling integration, exchange and reuse of its constituent parts. However, their framework has some weaknesses including limited querying and reasoning capabilities. Taylor et al. [128] presented a prototype for smart farming using Semantic Web standards to support real-time alerts for on-farm situation awareness. Their approach adopted SSN and other ontologies to represent knowledge of events over streaming data at runtime, publishing their summaries as linked open data. However, their research on enriching alerts with semantic linked data information is not complete.

Semantic annotations of sensor data enable applications to utilise enriched sensor data for different purposes including information exchange, reasoning, and creating context-aware applications etc. The work of Bernaghi et al. in [129], proposed a semantic data model to represent large heterogeneous data in a sensor network. They identified a major challenge in introducing semantics to sensor networks which is the addition of metadata to be exchanged alongside the measured data. However, their approach was based on O & M and SensorML specification which lacks explicit semantic interoperability. Wei and Barnaghi in [10], took the idea of semantic annotation a bit further and focused on using domain ontologies based on linked data principles. However, their work just advocated the idea and did not provide any details about the semantic enrichment process and data transformation to RDF using linked data principles. Broring et al. [130] presented a roadmap towards semantically

enabled sensor plug and play within the Semantic Sensor Web. Their approach focused mainly on semantic annotation of service requests which were made for adding new sensors and observations to the Sensor Observation Service. However, it lacked data integration and reasoning services. Huang and Javed in [131] proposed an architecture named SWASN, to describe and process sensor data to make it meaningful for other applications and to extract high level information from it. To demonstrate their work, they used a case study of a fire emergency scenario in a building. However, their approach lacked querying real-time data. Moreover, it was not scalable. Moraru and Mladenic [132] proposed a framework for enriching sensor data to improve its usability and accessibility. They built a semantic repository of sensor data containing both sensor descriptions and measurements that can be used by semantic browsers and inference engines. Their work provided a good conceptual framework, however did not provide any implementation.

**Analysis**

This section has surveyed Semantic Web technologies as they address some of the challenges in the IoT domain. There is a strong body of work in this area and hence a significant amount of experience and interest in applying Semantic Web technologies in this domain. However, all these research efforts have limitations in terms of fulfilment of the dimensions described in section 2.4.1. Firstly, the coverage of the ontologies is limited. For instance, some of the ontologies focus on data and measurements, with little mention of describing sensors, systems or measurement procedures, while others focus on sensors, systems and procedures but overlooking data and observations [133]. Hence, the coverage dimension is not fully satisfied. Secondly, some of the ontologies are not expressive enough. Though the SSN ontology is an important stepping stone and is relatively expressive, it needs to be extended and reused with other domain ontologies to provide a comprehensive solution for sensor networks. Thus, the expressiveness dimension is also partially addressed. Thirdly, a proportion of the work lacks semantic interoperability, data integration and querying support. This leads to the semantic gap for data interoperability, integration and querying dimensions. Finally, the semantic annotation and reasoning mechanisms are very basic and still require further research

| Reference | Purpose | Coverage | | | Expressiveness | Using Standards | Semantic Annotation of Data | Semantic Data Integration | Reasoning Over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| Avancha et al. | Adaptive Sensor Networks | * | ✓ | * | * | ✗ | ✗ | ✗ | * | ✗ | * |
| OntoSensor | Knowledge - base & Inference | ✓ | ✓ | * | * | * | ✗ | * | ✓ | * | * |
| Kim et al. | Sensors and Web Services | * | ✓ | * | * | * | ✗ | ✗ | ✗ | ✗ | * |
| SWAMO | Intelligent Agents | * | * | * | * | * | ✗ | * | * | * | * |
| A3ME | Low-powered Devices | ✓ | * | ✗ | * | * | ✗ | ✗ | ✗ | ✓ | ✗ |
| CSIRO | Integration, Search | ✓ | ✓ | * | * | * | * | * | * | * | ✓ |
| Sheth et al. | Semantic Sensor Web | ✓ | ✓ | * | * | * | ✓ | * | * | ✓ | * |
| Kuhn | Observation & Measurements | ✓ | ✓ | * | * | * | ✗ | * | ✗ | ✗ | ✗ |
| SSN | Sensors & Observations | ✓ | ✓ | ✗ | * | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Gray et al. | Semantic Sensor Web | ✓ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | ✓ |
| Wang et al. | Knowledge Representation | ✓ | * | * | * | ✓ | ✓ | ✓ | * | ✓ | ✓ |

| Reference | Purpose | Coverage | | | Expressiveness | Using Standards | Semantic Annotation | Semantic Data Integration | Reasoning over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| Barnaghi et al. | Computing Perceptions | ✓ | ✓ | * | * | ✓ | ✓ | * | * | * | ✗ |
| Roda and Musulin | Data Analysis on Sensor Measurements | ✓ | ✓ | * | ✓ | ✓ | ✓ | * | * | * | * |
| Taylor et al. | Web of Things | ✓ | ✓ | * | ✓ | ✓ | * | ✓ | * | * | ✓ |
| Barnaghi et al. | Semantic Model for Sensor Data | ✓ | ✓ | * | * | ✓ | ✓ | * | ✗ | * | * |
| Wei and Barnaghi | Semantic Sensor Web | ✓ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | * |
| Broring et al. | Semantic Sensor Plug & Play | ✓ | ✓ | * | * | ✓ | ✓ | * | * | * | * |
| Huang and Javed | Semantic Sensor Networks | ✓ | ✓ | * | ✓ | ✓ | ✓ | ✓ | * | * | * |
| Moraru and Mladenic | Semantic Enrichment of Sensor Data | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | * | * | * | * |

**Table 2.1:** The dimensions of the Related Work and their support in the survey of Semantic Web technologies for IoT/streaming data. The tick mark (✓) represents that the dimensions are fully satisfied, the cross symbol (✗) shows that the dimensions are not supported at all, and the asterisk symbol (*) shows the partial fulfilment of the dimensions.

to convert low-level sensor measurements into high-level knowledge. Hence, the dimensions of semantic annotation and reasoning are partially fulfilled. In short, there is a lot of good work and experience in this area, but most of the dimensions are not fully addressed. Hence, more research is needed especially developing a sensor ontology providing good coverage and reasoning capabilities and enabling integration and interoperability of different data sources effectively. The analysis is summarised in Table 2.1.

**(b) Semantic Web in Environmental Science**

Environmental data can play an important role in addressing the key challenges such as climate change, loss of biodiversity and sustainability of environmental ecosystem services to name but a few. As environmental science encompasses various other disciplines, it requires multidisciplinary collaboration and access to diverse data from interconnected sub-disciplines. In order to solve difficult research questions collaboratively, environmental scientists also need to access, use and share the data. Unfortunately, environmental data is usually stored in non-standardised formats, placed in geographically scattered locations and managed by different local, national and international authorities. These characteristics ultimately provide a hindrance to capturing complex interrelationships across datasets, wider data discovery and access, interoperability, data integration and reuse [134]. The Semantic Web offers the potential to introduce machine understandable semantic metadata with the help of ontologies and linked data mechanisms to address these challenges.

Researchers have developed controlled vocabularies, community thesauri and formal ontologies to potentially resolve the data challenges including discovery and access, data integration and interoperability. Controlled vocabularies and community thesauri can enable seamless description and presentation of data. They are used to characterise datasets and can be helpful in data discovery and integration process. This practice has been documented in, for instance, [135]. These approaches are a good starting point but they cannot provide rich and unambiguous semantics to infer new terms and knowledge.

Ontologies have been introduced to achieve precise and formal semantics. SWEET (Semantic Web for Earth and Environmental Terminology) [136], developed by

NASA Jet Propulsion Laboratories, is a set of more than 200 ontologies in the field of earth sciences. SWEET ontologies have been developed to improve the discovery and usage of earth sciences data through semantically enabled software. These ontologies conceptualise several categories of information including the earth realm, living and non-living elements, physical properties and spatio-temporal concepts. However, some of the SWEET concepts are interdependent within or across the ontologies and reusing or extending them would be overwhelming unless a structured approach is followed by the domain experts to analyse the gaps in the upper-level design [137]. Moreover, SWEET ontology represents broad information focussing on the taxonomy of domain specific events and provides fundamentally class hierarchies but limited expression of properties. The Extensible Observation Ontology (OBOE) is another approach that provides a semantically enabled metadata paradigm to facilitate discovery and interoperability of different geoscience datasets [138-139]. OBOE was used in the context of the Science Environment for Ecological Knowledge (SEEK) project that aimed at developing technologies (e.g. scientific workflows) for discovery, integration and analysis of distributed ecological data and information. Though the OBOE ontology model provides better interoperability, its reasoning performance is limited. Moreover, it does not support higher level context or constructs to describe a sequence of observations, e.g. in capturing an extreme weather event. The Network of Excellence project, ALTET-Net, developed the SERONTO (Socio-Ecological Research and Observation Ontology) ontology to integrate biodiversity data from distributed data sources [140]. SERONTO was tested through a biodiversity use-case; however it has some unsatisfiable concepts/classes which is fundamentally a modelling error leading to barriers in extending the ontology. Moreover, reasoning and inconsistency issues can arise because of these classes. In the field of biology and biomedical studies, the Environmental Ontology (ENVO) was developed to enable retrieval and integration of broader biological data [141]. The interesting feature of ENVO is the ability to annotate any environmental terms/components, however it mainly focuses on biological terms/data, and hence it cannot readily be used more widely in environmental science. Later, due to the growing need of environmental semantics, the authors attempted to extend the coverage of ENVO ontology to meet the requirements of other disciplines including ecology and biodiversity [142]. However, the extension raised other issues including ontology mapping, consistency etc.

A number of approaches based on ontologies and Semantic Web tools have also been adopted to address capturing complex interrelationships across datasets, data discovery, integration and interoperability challenges. Parekh et al. [143] proposed a semantic metadata management system using ontologies to address the data discovery issue and provide a basis for interoperability. The good aspect of their work is using existing domain ontologies including SWEET. However, the approach is not based on any standard temporal or spatial ontologies and does not support any reasoning or inferencing. Furthermore, the ontology has not yet been fully evaluated. The approaches described by [144] and Madin et al. [139], both based on the OBOE ontology, are examples of relatively better data discovery and integration techniques. These approaches provide better interoperability but are limited in terms of search and reasoning facilities. Berkely et.al in [17], presented a semantic search system and described how ontologies such as OBOE and formal reasoning can be exploited to enhance keyword search by applying semantic annotations in order to provide semantic descriptions of scientific observations. They extended the previous work on EML [145] and Madin et al. [139]. However, their approach does not support advanced search and data integration. The work of [146] introduced a semantic based approach, based on mark-up languages and domain ontologies, for integrating different geoscience datasets. As a proof of concept, they implemented a semantically enabled service oriented computational infrastructure called DIA (Discovery, Integration, Analysis) to support earth scientists to discover, analyse and integrate their data. Though it is a good research effort for data integration in geoscience, it suffered from performance issues with large datasets. [147] developed an approach, based on OBOE, to enhance the discovery and integration of heterogeneous ecological datasets. Extending the Ecological Metadata Language (EML) and supporting tools, they used semantic annotations to express and represent datasets with terms and vocabularies from domain specific ontologies. However, their approach provides a very preliminary form of data integration and does not involve reasoning mechanisms to provide compatibility of annotated measurements. This further leads to lacking support for automated data integration. The work of [148] applied data mining techniques in conjunction with an ontology of causation to help domain experts in identifying possible causal relationships between fish movement patterns and environmental drivers such as moon cycles, high river flow or high/low temperature. However, their ontology is a general conceptual model, which is not

based on formal axioms and reasoning, hence the approach lacks the reasoning capability.

Because of the variety of sub-disciplines (e.g. biology, ecology, hydrology, climatology, meteorology, oceanography and biodiversity), interdisciplinarity and collaborativeness in environmental science, data heterogeneity and integration issues occur [149]. Environmental scientists connected to these subfields use their own terminologies, different measurement units, different data models and experimental designs that exacerbate such data heterogeneity problems. To cope with the data heterogeneity challenge, the research community have provided some potential solutions through applying structured and standardised metadata approaches including standardised mark-up languages, for instance, the Ecological Metadata Language (EML) [150], the Earth Science Mark-up Language (ESML) [151], and the Water Mark-up Language (WML) [152]. However, these approaches cannot completely resolve the semantic interoperability issues. To overcome the limitations of these approaches, researchers have proposed the use of controlled vocabularies and ontologies to semantically integrate heterogeneous data, e.g. see [153] and [154]. The former approach benefited from using ontologies regarding heterogeneous data integration and querying and retrieval support. However, it lacks comprehensive reasoning and inferencing support. Besides, the approach is not fully evaluated. The good feature of the work in [154] is that it provided both more granular representation of environmental data and flexible methods of integration and querying. However, it suffers from scalability issue and becomes impractical for large amounts of data.

Linked data approaches are potentially useful in supporting data integration and interoperability by providing a homogeneous view of distributed data and making this view available for other researchers, e.g. see [155]. The contribution of the said approach is the integration of different ecological resources using linked data principles and the provision of reasoning capacity to infer new information from the stored data. However, the approach is based on neither any existing standards nor their own designed ontology, rather uses local data published in RDF, which is rewritten as an application ontology. Moreover, the reasoning capability of the approach is very rudimentary and is not comprehensive enough. The work of [156] also adopted the linked data approach to integrate and share ecological data stored in

underlying distributed databases. Exploiting linked data principles, they improved slightly data integration and sharing beyond the existing metadata capability with databases. However, their approach is not based on an associated ontology, hence suffered from drawbacks including insufficient descriptions of the datasets, difficulties in schema-level integration, and no support for reasoning capability. Shaon et al. [157] is an example of an open-source linked data framework for integrating and publishing heterogeneous geospatial data as linked data, developed under the UK Location Strategy [158]. The framework, developed by the GeoTOD-II project, implemented a set of draft guidelines which were released by the UK Cabinet Office for promoting and publishing geospatial linked data. The authors also intended to address the challenges associated with these guidelines, for instance, designing implementable URI sets for location, representing legacy geospatial data and developing ontologies for this data. The candidate framework was a good effort to provide a flexible means for integrating and publishing both current and new datasets in the linked data format. However, it does not use any existing standard ontologies, thus leading to semantic data integration issues. Their approach also lacks a developed mechanism for mapping geospatial data to RDF schema and ontologies, that can further create mapping problems. Moreover, the work is as yet not fully evaluated.

**Analysis**

In this section, the Semantic Web approaches have been surveyed which were proposed to address data challenges including capturing complex interrelationships across datasets, data integration and reasoning, interoperability, and data discovery in environmental science. There is an important body of work in this area and hence a considerable amount of effort in applying ontology-driven approaches in this domain. However, all these approaches have limitations in terms of satisfying the dimensions described in section 2.4.1. Firstly, there is a lack of domain ontologies to provide enough breadth to capture concepts across a range of sub-disciplines in environmental science. This leads to the partial fulfilment of the coverage dimension. Secondly, most of the approaches are not standardised. Hence, the dimension of using existing standards in not satisfied. Thirdly, the data integration and interoperability mechanisms for heterogeneous environmental datasets are still not well-established.

| Reference | Purpose | Coverage | | | Expressiveness | Using Standards | Semantic Annotation of Data | Semantic Data Integration | Reasoning Over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| Porter J. | Controlled Vocabulary | ✗ | * | * | * | * | * | * | * | * | * |
| SWEET | Earth Sciences Data Discovery | ✗ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | * |
| OBOE | Discovery and Interoperability | ✗ | * | * | * | * | * | ✓ | * | ✓ | * |
| SERONTO | Data Integration | ✗ | ✓ | * | * | * | * | ✓ | * | * | * |
| ENVO | Representation of Data | ✗ | * | * | * | ✓ | ✓ | ✓ | * | * | * |
| Parekh et al. | Semantic Metadata | ✗ | * | * | * | * | * | * | * | * | * |
| Berkely et al. | Semantic Search | ✗ | ✓ | * | * | * | ✓ | * | * | * | * |
| Malik et al. | Semantic Integration | ✗ | * | * | * | * | ✓ | ✓ | * | * | * |
| Leinfelder et al. | Semantic Integration | ✗ | ✓ | * | * | ✓ | ✓ | * | * | * | * |
| Bleisch et al. | Causal Relationships | * | ✓ | * | ✗ | ✗ | * | * | * | * | ✗ |

| Reference | Purpose | Scope | | | Expressiveness | Using Standards | Semantic Annotation | Semantic Data Integration | Reasoning over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| EML | Knowledge Representation | ✗ | ✓ | * | * | * | * | * | * | * | * |
| ESML | Addressing heterogeneity | ✗ | ✓ | * | * | * | * | * | * | * | * |
| WML | Hydrologic Information | ✗ | ✓ | * | * | * | * | * | * | * | * |
| Fox et al. | Semantic Integration | ✗ | ✓ | * | * | * | * | ✓ | * | * | * |
| Tarasova et al. | Semantic Integration | ✗ | ✓ | * | * | ✓ | * | ✓ | * | * | * |
| Moura et al. | Linked Data | ✗ | ✓ | * | * | * | * | ✓ | * | * | * |
| Mai et al. | Linked Data | ✗ | ✓ | * | * | * | * | * | * | * | * |
| Shaon et al. | Linked data | * | ✓ | * | * | * | * | * | * | * | * |

**Table 2.2:** The dimensions of the Related Work and their support in the survey of Semantic Web technologies in environmental science. The tick mark (✓) represents that the dimensions are fully satisfied, the cross symbol (✗) shows that the dimensions are not supported at all, and the asterisk symbol (*) shows the partial fulfilment of the dimensions.

Thus, semantic integration and interoperability dimensions are partially qualified. Fourthly, a large proportion of the work performs rudimentary reasoning and lacks comprehensive inference mechanisms to derive further new knowledge from the existing one. This leaves a semantic gap for holistic reasoning approaches and hence the dimension of semantic reasoning is not fully addressed. Finally, as compared to other disciplines, for instance, health care and life sciences, the uptake of Semantic Web technologies in environmental science is lower. Hence, further research is required to fill this semantic gap in understanding the highly complex and heterogeneous environmental data. In terms of this thesis, there is a particular need to further explore Semantic Web technologies to understand the characteristics of environmental science around data. The analysis is summarised in Table 2.2.

**(c) The Semantics Web for IoT/Streaming Data in Supporting Environmental Science**

As discussed above, the research in this thesis sits at the intersection of all three areas, i.e. the Semantic Web, IoT/streaming data and environmental science (**Figure 2.4**, marked in red). The related work in this area is summarised below.

Although there has been quite a lot of research on ontologies for sensor networks (as discussed in section 2.4.1 (a) above), there is very little research specifically targeting environmental science. There exist a few ontologies for IoT/streaming data in supporting environmental science. In oceanography, the Marine Metadata Interoperability (MMI) ontology was developed to describe oceanographic devices, including both sensors and samplers [159]. The ontology specified system concepts, its components and organisation of these components. MMI was used to enable users or applications to discover sensors and exchange and integrate marine data. This is an interesting initiative but the work is relatively immature in terms of development or evaluation.

The Coastal Environment Sensor Network (CESN) project designed and developed an ontology [160] as part of the Semantic Data Reasoner project for coastal observation to infer ecosystem events. The ontology was built to encode sensor types and was based on Description Logic and logic rules to deduce inferences about sensor data and also detect anomalies. The strength of CESN lies in covering a wide range of

ontology concepts and the capability of reasoning domain knowledge from data. However, the project encountered knowledge modelling issues including an excess number of classes, which limited the scalability of the model. Another issue was conflating observation data with the properties of sensors potentially leading to semantic data integration issues.

The work of [161] described the AEMET (Agencia Estatal de Meteorologia) ontology network, developed for meteorological forecasting by the Spanish meteorological bureau, to transform the meteorological data into linked data. The goal of the approach was to describe sensor measurements, generated by the network of meteorological stations. The AEMET ontology also reused the SSN ontology. The good feature of the AEMET ontology is that it is a modular ontology that describes time and location concepts in addition to sensors and measurements. However, as the approach of [161] was performed in parallel to the development of SSN ontology, some of the design decisions of the approach for transforming meteorological data are not completely compliant with the existing SSN ontology.

Once the sensor data is enriched with semantics, it can help ontology to reason over it and deduce new knowledge from it. The work of [119] reasoned over heterogeneous data to infer a blizzard event. In a similar approach, Wei and Barnaghi [10] performed rule based reasoning over semantically enriched sensor data to derive the condition of 'potentially icy' road. Henson et al. [162] proposed an ontological model of time series observations to add value to sensor data on the Semantic Sensor Web. Using rule based reasoning over sensor data, they specified weather events in the environment including 'blizzard'. Devaraju and Kauppinen [163] developed an ontology and reused the DOLCE ontology to capture different weather properties and investigate how blizzard events can be inferred in regard to observed atmospheric properties. However, their approach used only upper ontologies with no other ontology to specify sensors and measurements. Su et al. [164] proposed an approach for reasoning over sensor measurements by taking a use-case from the fishery IoT system to deduce alerts and reminders. In [165], Thirunarayan et al. illustrated to represent and enhance raw sensor data with spatial, temporal and thematic annotations to detect inconsistent sensor data. Their approach formalised data from the Weather domain and reasoned over it using a meta-interpreter in Prolog. To summarise, all the

above semantically-enabled inference approaches were interesting initiatives of reasoning over sensor data to deduce new knowledge. However, they performed very preliminary reasoning and none of them provided a holistic approach to spatio-temporal inference of knowledge or events. Furthermore, these approaches were not based on a standard spatial and temporal ontology having controlled terms to describe either complex interval-based temporal or spatial events and perform reasoning over spatio-temporal operators.

Yu and Taylor in [166], proposed the Event Dashboard, a web based user application capturing semantics for events of interest in a sensor network. The Event Dashboard provides an ontology-driven user interface for detection of algal bloom events over sensor data in a sensor network. The authors aimed at resolving the data heterogeneity issue of sensor networks by using a domain ontology. Their work extended the SSN ontology and used a case study in the water quality domain to model observations around the chemical properties of water. This work is a good initiative to enable users to express event constraints using the SSN ontology, however the drawback of their approach is both the high degree of complexity that lies in the underlying set of ontologies driving the user interface (UI) and an overhead over defining queries in an event processing engine.

Roussey et al. [167] described the process of publishing RDF datasets from meteorological stations. Their work aimed at reusing existing standards and tools. This work was a good example of using existing ontologies but the work did not provide any new insights or methodologies for this area. Lefort et al. [168] described a similar approach of transforming and publishing ACORN-SAT climate data as linked data. They captured and integrated their temperature time series datasets using the SSN ontology and published them as linked data. The publication of ACORN-SAT datasets is the first initiative of linked data published by the Australian government. However, their approach lacks reasoning and deducing new knowledge.

**Analysis**

As can be seen from the work above, there is an interest in the use of Semantic Web technologies for IoT/streaming data and supporting environmental science. However, the state-of-the-art is still limited in terms of fulfilment of the dimensions described in section 2.4.1. The analysis is summarised here. i) The ontologies do not provide

| Reference | Purpose | Coverage | | | Expressiveness | Using Standards | Semantic Annotation of Data | Semantic Data Integration | Reasoning Over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| MMI | Marin Metadata Interoperability | ✓ | ✓ | * | * | * | * | ✓ | * | * | * |
| CESN | Knowledge Inference | ✓ | ✓ | * | * | * | * | * | ✓ | * | * |
| AEMET | Linked Data | ✓ | ✓ | * | ✓ | ✓ | * | * | * | * | * |
| Sheth et al. | Semantic Sensor Web | ✓ | ✓ | * | * | * | ✓ | * | * | * | * |
| Wei and Barnaghi | Semantic annotation and Reasoning | ✓ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | * |
| Henson et al. | Reasoning over Sensor Data | ✓ | ✓ | * | * | * | * | * | * | * | * |
| Devaraju and Kauppinen | Reasoning over Sensor Data | ✓ | ✓ | * | * | * | ✓ | * | * | * | * |
| Su et al. | Transforming SenML to RDF | ✓ | ✓ | * | * | * | * | * | * | * | * |
| Thirunarayan et al. | Sensors & Observations | ✓ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | * |
| Yu and Taylor | Capturing Semantics for Events | ✓ | ✓ | * | * | ✓ | ✓ | ✓ | * | * | * |

| Reference | Purpose | Coverage | | | Expressiveness | Using Standards | Semantic Annotation | Semantic Data Integration | Reasoning over Data | Data Interoperability | Querying Support |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sensor | Observation | Domain | | | | | | | |
| Roussey et al. | Publishing Meteorological Data | ✓ | ✓ | * | * | ✓ | ✓ | * | * | * | * |
| Lefort et al. | Transforming Climate Data | ✓ | ✓ | * | ✓ | ✓ | ✓ | * | * | * | * |

**Table 2.3:** The dimensions of the Related Work and their support in the survey of Semantic Web technologies for IoT/streaming data in supporting environmental science. The tick mark (✓) represents that the dimensions are fully satisfied, the cross symbol (✗) shows that the dimensions are not supported at all, and the asterisk symbol (*) shows the partial fulfilment of the dimensions.

enough coverage to support both the IoT and environmental science domains. This leads to the partial fulfilment of the dimension of ontology coverage. ii) The above work provides reasoning either at a very basic level and/or has limitations in terms of their support for ontological reasoning and therefore does not provide a comprehensive solution for drawing inferences over environmental sensor data. Thus, the dimension of semantic reasoning is not fully addressed. iii) Some of the proposed solutions focus on providing interoperability between sensors instead of the (higher level) data collected from the sensors. Hence, the dimension of semantic interoperability is partially satisfied. iv) Some approaches impose limitations on querying support while others on heterogeneous data integration and interoperability. v) There is less research on integrating, reasoning and querying real heterogeneous data from sensor networks deployed in the natural environment. Hence, the uptake of Semantic Web technologies for IoT/streaming data in supporting environmental science is not fully realised, leaving this dimension partially addressed. The analysis is summarised in Table 2.3.

## 2.5 Summary

As mentioned in the introduction, the aims of the chapter were twofold: to give an overview of technological developments and to examine the state-of-the-art in the Semantic Web for environmental science. Therefore, we reviewed the background knowledge, placing this work in context, by offering a broader perspective on eScience and one of its enabling underlying technologies, underpinning environmental science. Then, we surveyed the state-of-the-art in the areas of the Semantic Web, IoT/streaming data and environmental science, in accordance with the research goals mentioned in Chapter 1. To summarise, the chapter concludes with the following key points:

- Work to date remains relatively tightly focused on single dimensions of the environment, lacking a broader view that can integrate and reason over data across multiple scientific sub-domains to build a holistic environmental perspective.
- A large proportion of the above work is technology-oriented and often fails to study the emerging trends and events stemming from the real environmental data.

- A body of the related work in sensor networks focuses on underlying networking technologies, sensor discovery mechanisms and the development of services but there is less research on interoperability, data integration and reasoning, and querying real heterogeneous data from sensor networks deployed in the natural environment.
- The state-of-the-art has limitations in terms of their support for ontological reasoning and hence do not provide a comprehensive solution for drawing inferences to deduce new knowledge.
- In environmental science, there is a lack of both sensor and domain ontologies that can provide enough breadth to capture thematic, spatial and temporal dimensions of environmental data across a range of sub-disciplines.
- The uptake of Semantic Web technologies in the context of IoT/streaming data underpinning environmental science is low and examines mostly single facets of the natural environment.

From the analysis, we further conclude that all of the above work suggests a strong need for further exploration of Semantic Web technologies and associated techniques. In contrast to these points, there is a need for research to take a multi-dimensional perspective on environmental IoT data understanding it in all its complexity. Hence, further research is required to apply Semantic Web technologies allowing new scientific insights to be gained through examining environmental data in novel ways.

The next chapter further examines the characteristics of environmental science around data, through semi-structured interviews, to develop further the research questions and surrounding perspectives for this thesis.

# 3 Qualitative Study of Data Challenges in Environmental Science: Understanding the Long Tail of Science

This chapter reports on a qualitative study of environmental data and shares insights gained from the interviews with leading environmental scientists. More specifically, the main goal of the chapter is to examine the unique characteristics of environmental science in the context of environmental data, through semi-structured in-depth interviews. This goal can be further divided into the following more specific objectives:

- Learning how embracing open data approach can bring benefits to environmental science and further enhance interdisciplinary and multi-disciplinary collaboration between environmental scientists.
- Gaining knowledge and understanding of the data needs, limitations, frustrations and technological barriers the environmental scientists are facing.
- Achieving new academic understanding of discovering data and the interdependencies across disparate datasets.

- Informing how to develop a generic semantic data model of integrative environmental science, allowing the examination of different environmental datasets in novel ways.

This chapter investigates: the role of data and data practices in environmental science; a potential paradigm shift in science towards open data; interdependency between disparate datasets representing different environmental facets; technological opportunities the environmental scientists gain from collaboration with computer scientists and engineers; and barriers such as data discovery and access, data heterogeneity, data integration and interoperability. The overarching purpose is to provide deeper understanding of the area to inform the approach developed in the thesis in terms of applying Semantic Web technologies to the natural environment, both in terms of the overall context of trends in data needs, and also in terms of identifying specific requirements.

# 3.1 Methods

The methodology adopted is a mixed methods approach based on semi-structured in-depth interviews coupled with a Grounded Theory approach [169] to extract insights and meaning from the resultant transcripts.

This study has been assessed and approved by the Research Ethics Office, Lancaster University. We provided the consent form to all participants which they returned after signing them. All their information has been treated with confidentiality. They have the right to withdraw permission from the study within two months of data collection, and if required, to have their data collected withdrawn from the study. If they withdraw before the deadline, their data will be destroyed and will not be used or remain in the study but if they do after the deadline, their data will remain in the study. Data will be stored in ways to make sure their identity cannot be inferred.

The method is discussed and justified in detail below.

## 3.1.1 Semi-structured In-depth Interviews

This work has been done in close collaboration with environmental scientists who own and use rich environmental data. A series of semi-structured in-depth interviews

have been conducted with environmental scientists from three different universities in the UK and the Centre of Ecology and Hydrology (CEH). Semi-structured interviews have been used for the following reasons.

- This approach supports predetermined but open-ended questions in order to allow a fair degree of freedom and flexibility, allowing new questions to emerge from the dialogues.
- Semi-structured interviews allow the interviewer to delve deeply into the topics so that detailed knowledge of the domain is gained.
- This technique keeps the interview focused, conversational and allowing two-way communication.

The domain experts (interviewees) have been chosen due to their experience and considerable expertise in their disciplines. They are not data naïve, but rather have already transitioned into data science and have been using data in a sophisticated way. Furthermore, they are at the forefront of data-driven environmental research and bring the sort of broad and holistic perspective of environmental data the thesis is looking for.

The domain experts spanned a wide range of environmental science including ecology, hydrology, soil science, environmental chemistry, volcanology, climatology, molecular and microbiology, limnology and meteorology. All these interconnected sub-disciplines have been chosen to get an integrative understanding of different environmental datasets, how they are related to each other and how development in one discipline can impact the other.

A total of 18 semi-structured interviews were carried out and, at that point, it was determined that saturation had been achieved [170], and hence no further interviews were deemed necessary.

The interviews were planned to contain a number of questions covering five categories, i.e. data role and practices, trends in data management including openness, collaboration, and integration, focus on interdependency between disparate datasets, technological opportunities and technological barriers. These broad areas are not arbitrary, but rather have been extracted from the author's reading and understanding

of the literature as described in Chapter 2 and they make a comprehensive set of fundamental top-level issues around data management in environmental science. Interviews ranged from 50 minutes to one and half hours with an average of one hour per interview. All interviews were audio recorded and later transcribed. The interview questions, classified into five categories, are shown in Box **3.1**. However, the interviews are not restricted to only these sets of questions following the semi-structured approach.

### 3.1.2 Grounded Theory

To interpret and analyse these in-depth interviews, a grounded theory approach was used [169]. Grounded theory is defined by Strauss and Corbin as a qualitative research methodology for developing theory that is inductively grounded in data systematically gathered and analysed [171-172]. The evolution of theory occurs during actual research through continuous interaction between analysis and data collection which is the key feature of this analytical approach referred to as constant comparative method [172-173]. The grounded theory approach consists of several analytical steps that are non-linear and recursive. The steps in this research are based on the works of Glaser's [174], Charmaz [175], Chesler [176], and Strauss and Corbin [171] analytical method of theory development that are shown in Figure 3.1 and Figure 3.2. These steps are described below.

The first step is about collecting data that was captured from the semi-structured interviews (section 3.1.1). After collecting and examining a rich set of data (resultant transcripts) from the interviews, the data was broken down into discrete chunks and coding was performed which is the key part of grounded theory methodology. Codes are shorthand devices that are used to label and organise the data [175]. The author highlighted key phrases in the data and assigned different codes to those key terms [176]. For instance, the participant's data, *"Data is the 'lifeblood' of climate science and is central to understand atmospheric composition and climate change*" is assigned the code "*data is the lifeblood of environmental science*". Similar code phrases were grouped together to be reduced and then organised into clusters. Clusters were reduced and labels were attached to them. These labels are called concepts. Similar concepts were grouped together to form categories (classification of concepts). Glaser and Strauss' [173] method of constant comparison was performed to

compare codes and categories for similarities and to identify different categories and reflect on different potential relationships across categories. These categories and concepts were interlinked and core categories were identified which are the central themes of the data [171,175]. From these core categories, observations (cf. mini-theories) were generalised which led to the emergence of overarching themes about the unique characteristics of data in environmental science.

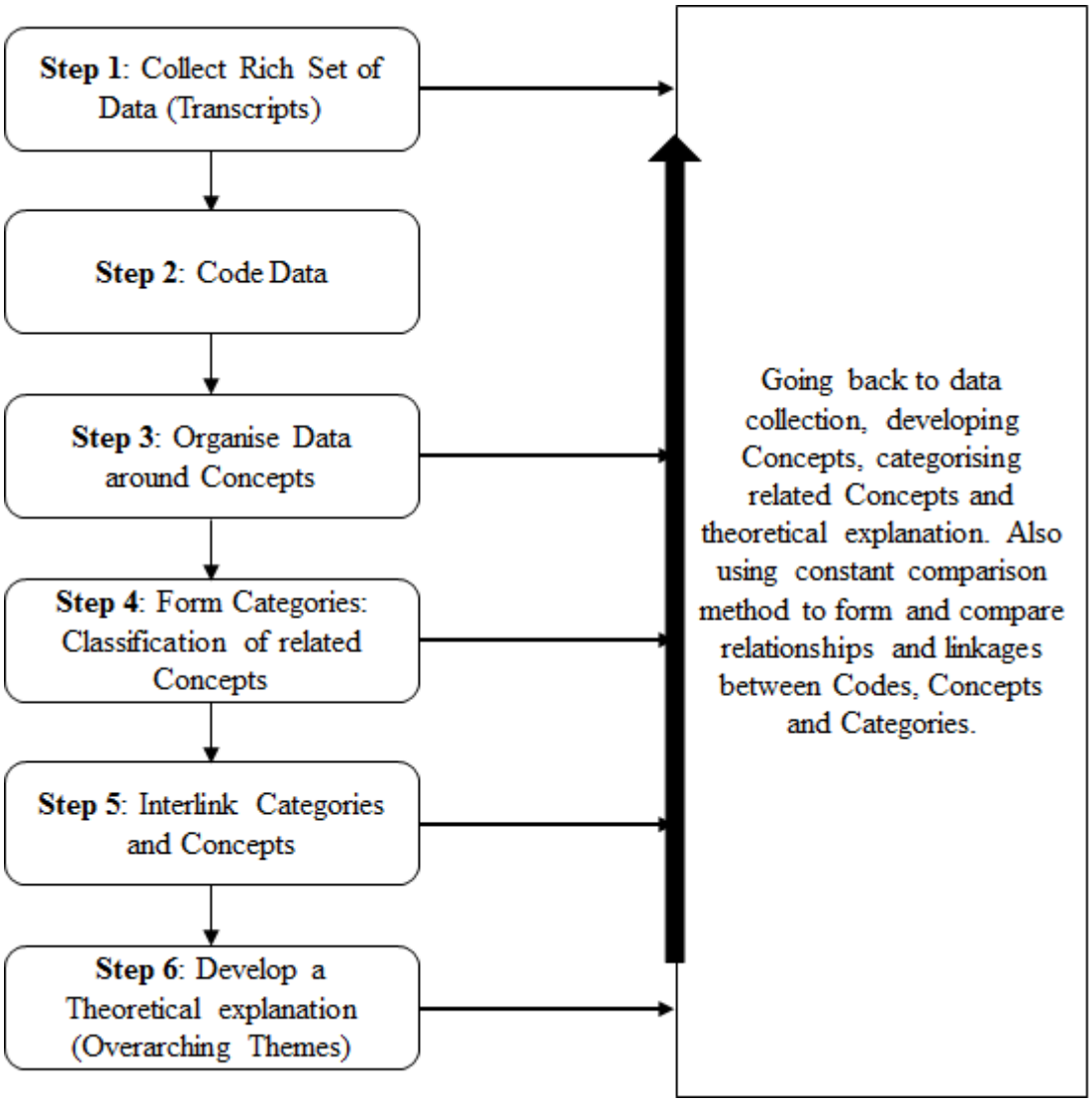


**Figure 3.1 Ground Theory Analysis based on the work of Glaser [174]**

The detailed diagram describing grounded theory analytical method based on the works of Charmaz [175], Chesler [176], and Strauss and Corbin [171] is given in Figure 3.2.

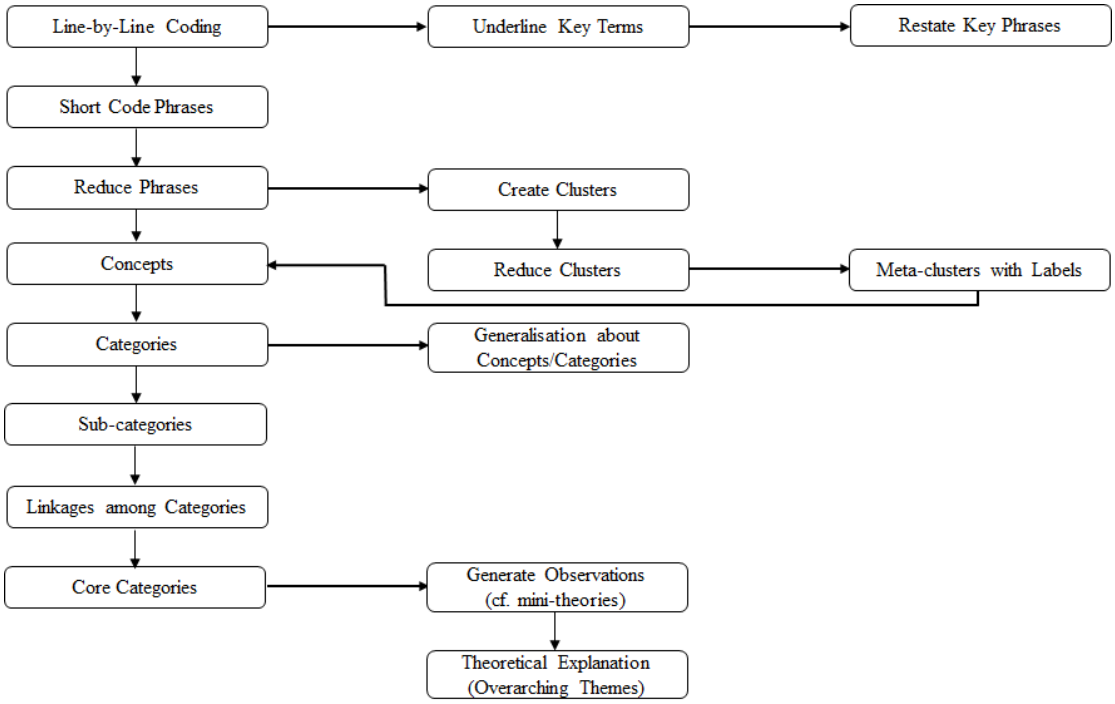


**Figure 3.2 Ground Theory Data Analysis based on Charmaz [1983], Chesler [1987], and Strauss and Corbin [1990]**

Figure 3.2 Ground Theory Data Analysis based on Charmaz [175], Chesler [176], and Strauss and Corbin [171]

A sample of emergent core category, sub-category and initial codes are given below.

| Core Category | Sub-category | Codes |
|---|---|---|
| The Importance of Interdependency | Causal-like Relations | Impact of land management on soil quality, relationship between soil and water quality, the link between atmospheric science and soil science, link between atmospheric composition and climate change |

The results of the interviews are organised into five different sections on the basis of research questions in Box 3.1. The author presents findings from these interviews and then reflects on overall messages with respect to the context of this thesis.

## Box 3.1: Interview Questions

**The Role, acquisition and storage of Data**

- What is the role of data in your science?
- What practices and technologies do you currently use to capture your data?
- How do you go about storing data?

**Trends in Data management: Openness, collaboration and integration**

- Do you personally offer open access to your data and if not, why not?
- What problems do you face in open data approach?
- Do you think an open data approach can bring benefits to environmental science generally?
- Do you see open data as being a focal point to enhance collaboration between environmental scientists?
- Is this something you currently do and, if not, why not?
- How important is the integration of datasets in your work?
- Do you see this as becoming more or less important in the future?

**Interdependencies in the Long Tail of Environmental Science**

- What other kinds of data would you like?
- Have you heard about the long tail of science and to what extent this applies to your work?
- More specifically, how important is inter-dependency in your work, e.g. identifying causal-like relationships between datasets (could you provide example)?
- When you work with data, do you typically take a positivist approach, seeking to prove or disprove a hypothesis, or do you see room for more emergent approaches?

**Technology: Opportunities**

- Do you see collaboration with computer scientists is important in your work and, if so, what would you like to gain from this?
- What are the potential barriers to collaboration with computer scientists?
- Is this something you currently do, and what benefits have you got from this?
- How important is it generally for you to have a unified view of the structure and semantics of heterogeneous datasets?
- Which of the following techniques are you aware of, and which ones do you see as potentially contributing to your work in the future:

  Semantic Web Technologies (Ontologies, Linked Data), Statistical Methods, Data Mining and Machine Learning.

**Technology: Barriers**

- To what extent are the following real barriers in your work?

  Data discovery and access, problems with the quality of data, the heterogeneity of data sets, the lack of metadata or provenance information around data.
- What other technological barriers or frustrations do you face in your work as an environmental scientist, particularly around data?
- What single technological advance would you wish for (and you are encouraged to think big here), that would support you as an environmental scientist in the science you would like to do over the next 10 years?

## 3.2 The Role, Acquisition and Storage of Data in Environmental Science

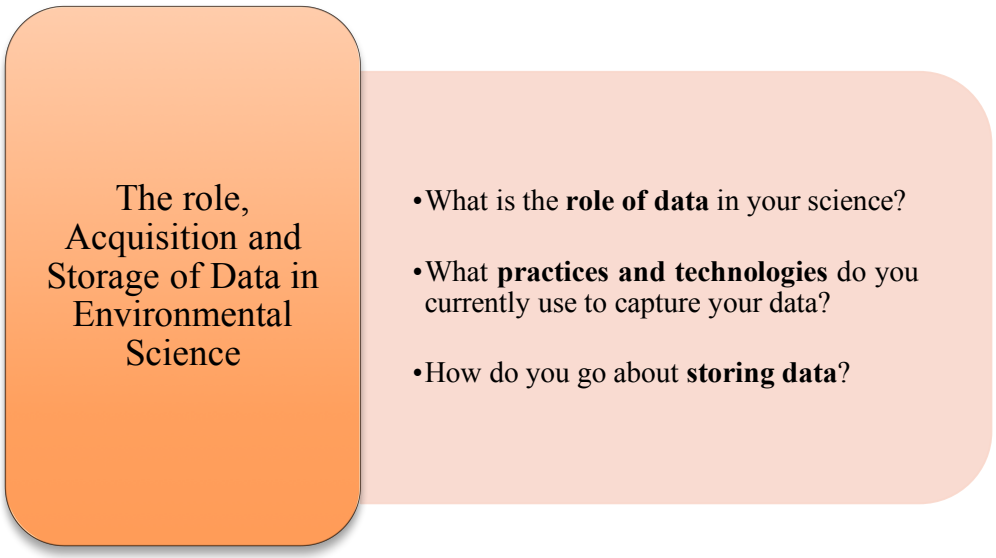


**Figure 3.3 Data Role and Practices**

### 3.2.1 Background

Data is critically important to understand and predict global environmental changes as well as the impacts of these changes. Environmental scientists use and analyse such data to address different challenges including among others loss of biodiversity, climate change and also to inform policy design and decision-making. In short, environmental data is required to understand and manage overall ecosystems. Hence, it is essential to archive environmental data, which involves acquisition, storage and preservation of data. In this section, findings have been drawn from the interviews by selecting responses to questions to gain insights into the role of data in environmental science, and the data practices and technologies environmental scientists use for data acquisition and storage, as shown in Figure 3.3.

### 3.2.2 Main Findings

One of the first main themes that emerged is the criticality of data across all the areas of environmental science under consideration. One of the participants, working in the

field of sustainable land use systems, summarises how data plays a key role as evidence for decisions:

*"Data is hugely important in my science because a lot of work I have done and still doing is to provide evidence for policy makers, for instance, in the management of livestock manures and other organic resources to optimise nutrient utilisation while minimising impacts on water and air quality. This includes mitigating greenhouse gas emissions, reducing the risk of transfers of pollutants such as nutrients and pathogens to watercourses, and understanding the secondary impacts of mitigating diffuse agricultural pollution. So, we collect data about these variables which are then used for validating models and calibrating models… whether they are mechanistic models or whether they are statistical models to try and explain some of the variability we see."*

Another participant, working in the development of long-term ecological research networks, explains how important data in their science is:

*"Data is absolutely essential because we work in an environment where you have to provide evidence for decisions, and evidence primarily comes from data. Data is often collected in the scientific field using some measurement technologies old or new which can be summarised into information and knowledge which feeds into the evidence process. You can't really have evidence without data backing it up some way."*

A climate scientist identifies the role of data in climatology:

*"Data is the 'lifeblood' of climate science and is central to understand atmospheric composition and climate change, and links between the two. I use computer models and simulations and use the output data to work on stratospheric and tropospheric ozone, multi-climate models' analyses, biosphere-atmospheric links, analysing temperature data and modelling novel pollutants. So, I can't pursue my research without data."*

A soil scientist recognises the essential role of data in her science:

*"We can't do anything without data. We mainly use data for quantifying environmental responses to evaluate whether there is environmental change essentially, for instance, understanding below-ground processes with specific focus on nutrients and human pathogen behaviour in soil-plant-microbial systems. So, what we do is quantitative*

*science where you need to have data. Then you need to analyse it in a way that you can see the variability in different responses from the environmental stimulants. So, we use data in statistical analyses to understand, for example, the use of wastes for land restoration, controlling bacteria such as E. coli in agricultural systems, enhancing food safety, carbon sequestration in grasslands and ways to improve nutrient use efficiency in cropping systems."*

The following environmental chemist who has a deep interest in how synthetic organic chemicals behave in the environment talks about the significance of data in his research:

*"The role of data is to make sense of environmental processes that govern the fate of pollution, for instance, industrial chemicals, pesticides and pharmaceuticals and those factors that affect their longevity in the environment, including in remote regions like the Arctic. I do contribute data to international programs so one example that I work with is the Arctic Monitoring Assessment Programme (AMAP) so this is looking at the industrial and agrochemical pollutants that have been washed or wafted into the arctic. The AMAP tries to bring together all the datasets that are being generated and provide an assessment report every three or four years. So, my data will go into that assessment."*

One of the participants, working as a hydrologist, describes the role of data in hydrological modelling and decision making for water management under uncertainty:

*"I am interested in data both observables, different types of observables as input to models and also as a constraint on uncertainty in models after we get some output from the models. We do a very large number of runs of models to try and investigate the uncertainties in the outputs. So, in part, my interest is how you put models' output and observe a boost together and in particular when because of time and space scales variables have the same names where they actually mean different things both as parameters in models and the outputs from models what you could actually observe, so soil moisture is a good example."*

The data used by all these participants has been playing an essential role in understanding and managing the environment such as benefitting human life for their

welfare and safety, reducing human losses caused by natural and anthropogenic calamities, helping in responding to climate change and its implications, better management of the ecosystem, protecting water resources from pollutants, enhancing the agriculture and conserving biodiversity. Some participants use their data to examine and monitor the meteorological and climatological variables to understand the climate system functions and predict the future. Using their analytical skills, they process and analyse these datasets to get knowledge of underlying - physical, chemical and biological processes. A few participants are undertaking research based on their data for the management and control of atmospheric and water pollution which is threatening human and animal health, vegetation and the overall ecosystems. They analyse their data to understand, assess and reduce risks posed by the organic chemicals and to examine critically the interaction between environmental pollutants and local communities. Some of the participants, working as soil scientists, use their data to understand plant-soil-microbe interactions, soil quality and how different environmental pollutants such as nutrients, pathogens, and sediments can affect the water quality and aquatic life. A couple of participants use their data to deal with extreme events such as flooding and predicting ecohydrological responses to future changes in catchments. They work with sparse datasets that may be subject to epistemic rather than aleatory error and uncertainties [177]. Two participants working in the area of environmental risk management evaluate the outputs of some very large ensembles of potential model representations as hypotheses in reproducing the characteristics of the test data, while allowing for potential uncertainties. In summary, environmental data is used by all participants not only for their own research but it also helps resource managers (water, land, health and marine resources) and policy makers to shape their decisions and develop strategies about environmental change respectively.

**Practices and Technologies**

Data in environmental science is acquired to produce and validate research results. Most of the participants use their own data collected from field observations through environmental sampling. They go out to the field site, collect a sample of some chemical, physical or biological phenomena (e.g. water, soil, plants, carbon flux, air, species, rainfall or temperature) using different sampling techniques and bring the

sample back to the laboratory for investigation and data analysis. One of the participants explained this practice:

*"We go out to the catchment area and collect soil samples using different sampling techniques and bring them back to the laboratory for experimentation and analysis."*

Environmental scientists use various ways to acquire the data. One of the participants, a climatologist, collects most of his research data from the Centre for Environmental Data Archival (CEDA) archive:

*"I download the data from CEDA archive with a file transfer program such as lftp script and then process the files using a mixture of Unix (bash) scripting and scientific software (NCO, Ferret and IDL). Climate model output is stored in a format called netCDF, and I generally convert observations to the same format, if they are not in that form already."*

The next set of comments highlights the increasing variety of sources of data. Sometimes environmental scientists engage citizen scientists for data collection because it is relatively a cost-effective way to acquire environmental data over large spatio-temporal scales. According to [178] citizen science is the *"volunteer collection of biodiversity and environmental information which contributes to expanding our knowledge of the natural environment, including biological monitoring and the collection or interpretation of environmental observations."* One of the participants below explains the role of citizen scientists in their data collection process:

*"We have other systems within our organisation that make a lot of useful field observations, for instance, species compositions, plants, water quality and structure in the landscape, from citizens; generally, we call them citizen scientists. These are experienced people that know and can identify plant species. They send a lot of records of particular species and contribute something to our centre called biological record centre which is based at the other branch of our organisation."*

The combination of low cost miniaturised embedded microprocessors, advanced sensing hardware, improved networking and communication technology and sophisticated data integration software have enabled environmental scientists to measure and monitor environmental variables over temporal and spatial scales which were impossible or expensive before [20]. The following ecologist, being part of a

project using sensor networks for measurements, describes entering in the era of sensor networks for data collection in their organisation:

*"And finally, we are moving into the world of sensor networks where we have an increasing use of sensors for measuring different facets of the environment. The sensors themselves are not necessarily new but the way they are deployed and the way data are being telemeterised to become the real-time picture of what's going on is an increasingly important part of the work we do."*

The following participant, serving as an environmental chemist, identifies the problem with his new automated sampling techniques:

*"We developed some automated novel sampling techniques but they won't actually give us the raw chemical data, those samples still had to be brought back to the laboratory to actually measure the chemicals which should have been captured by those samplers. So, it's quite a laborious technique. We can't just put out some automated instruments in the field that generate numbers and then by telemetry it sends packets of data back to me. No – I have got to actually take the sample, perhaps concentrate them in the field in some way and then take some media that might be a filter paper or whatever I use to capture that aspect of the environment, bring it back to the laboratory and then undertake chemical analysis of that media."*

Sometimes one technology might be useful for acquiring data in one environmental area but not in other scenario. One of the participants in the hydrology discipline illustrates this point:

*"Many of these remote sensing techniques are limited for hydrologists because they only review what's happening out in the first few centimetres of the soil and of course most of our interest happens at the greater depth than that. So, remote sensing technology in hydrology has not been useful yet, though it has the potential and promise that it would be more useful in the future. It could only be useful if you are working at global scales, then of course remote sensing is the only information you have to work from and so people do and use vegetation map, soil map and geological information from remote sensing."*

In summary, participants' data can be classified into different categories which include: observational data (including spatio-temporal measurements from various

sources e.g. field observations, weather station readings, satellite data etc.), experimental data (usually generated in a controlled or semi-controlled environment e.g. greenhouse experiments, chemical analysis etc.), simulation data (generated from models e.g. climate models), and derived data which is usually generated from other data files. Scientists acquire data through a number of methods and instruments such as hand-written notes, tape recorders, digital cameras, smart phones, laser scanning, close-range photogrammetry, mm-wave radar, infrared and remote time-lapse imaging, UAVs, data loggers, environmental sensors and satellites platforms. A couple of participants collect their data from the existing freely available electronic databases or archives. One of the hydrologists said that his subject area is lacking advanced measurement techniques, though there are a lot of theories around but most of them are not very good. They are waiting for new measurement techniques that become available in particular on large scale. As one of the hydrologists said, *"I made the argument, for example, that if we had the measurement technique that would measure the total storage in the profile at sort of 100 metre scale then we'd have different models and theories but that technique doesn't exist. Well, there is a technique using gravity anomaly but it's very expensive, takes a lot of maintenance and couldn't be widely applied. If somebody took, say the gravity anomaly technique and made that cheap which could be widely applied and easily maintainable that would revolutionise my subject area. So, in the future, I'd like to have new measurement technique but I've no idea what they can look like. Nothing is going to change very much in hydrology until new measurement technique comes along.*

In spite of advancements in automated measurements and environmental sensor technologies, some participants prefer to stay with their own reliable and easy to handle manual data acquiring and measurements techniques, as one of them explained, *"Well, I don't say we don't need advanced environmental sensing technology but they bring a lot of issues with them, for instance, increased complexity, reduced reliability, low trust in accuracy of data and sometimes they do not serve our purpose appropriately.* However, to measure and monitor the complex environmental phenomena appropriately, which change drastically over spatio-temporal scales, most of the participants recognise the need and importance of advanced automated technologies because of the methodological limitations in their current measurement techniques. They raised a valuable point that there is a need of increased multidisciplinary collaboration between environmental scientists, computer scientists and engineers to

contribute in the improvement of advanced sensing technology which could transform and expand the field of environmental science. If the technological challenges posed by these advanced sensor networks including energy efficiency, appropriate communication protocols, QA/QC, real-time data management and analysis are overcome, the measurements and monitoring in environmental science can be extended over larger spatio-temporal scales.

**Data Storage**

Data storage is a really important part of data-driven research and an important prerequisite to data sharing. If proper storage mechanisms and policies are not adopted, this may lead to the phenomenon of data decay and might further lead to less or no accessibility over time. In order to avoid this situation, best practices of data storage and management are required. Most of the data in the long tail of environmental science is collected either through hand-sampling methods or using some automated instruments. The data is recorded either in structured form such as database tables and spreadsheets, in semi-structured form such as XML files or in an unstructured form like plain text, images, sounds, videos and blogs etc. Asking about storage methods, one of the participants storing his data on portable devices said, *"I just store my data on regular portable storage devices and it's not too excessive. I guess all my data would be around one Gigabytes or something like that but it's not huge volumes".*

The participant below explains how their data is stored in their new project:

*"Well, usually I store most of my data on hard drive of my PC, flash drives and laptop disk but in this project, we just secured some additional funding to get all of our data into the right format that can then be uploaded and will be uploaded onto a data portal or data archive. So, we are planning to buy some additional hard drives and server machines to provide data backups for long term use."*

One of the participants, working in the data centre group, stores most of the scientists' data on their proper data servers using different types of database software (DBMSs):

*"We store our data in a variety of different formats and different infrastructures. So, spatial data will probably go into ArcInfo spatial database, NetCDF files are stored in a threaded data store called threads or gridded data store called threads I should say. And*

*then of course we do handle spreadsheets, image files, Matlab files and anything you can think of. So, we tend to store that information into non-proprietary formats and will turn into something like csv files. In short, we store our data in PostgreSQL databases, Oracle databases, gridded databases and spatial databases like ArcInfo etc."*

One of the participants, working at the data centre of a public-sector organisation, identifies the significance of data preservation for long term:

*"We have a data systems group where not only we store our data on different servers but we also make sure that our data are stored in a correct and consistent way along with their backups, they are safe over a long period of time I'll say an infinite period of time. So, it's hugely important that the data are 100 percent secure for a long term so that it could be (re)used by other scientists to promote new research and investigation in science."*

To conclude, most of the participants store their data on laptops, external hard drives, USB sticks, CD ROMs. Few of them use institutional data servers or centres. Most of the datasets they collect are usually small and heterogeneous. After data collection, the participants manage and organise their data using a number of applications and software including Microsoft Excel and MS Word, scripting languages such as R, Matlab and Python, statistical packages including SPSS, SAS and some database software e.g. Oracle DBMS, Microsoft SQL Server, PostgreSQL and MySQL. Most of the participants are facing challenges of persistent storage, data curation and preservation because often they do not get funding for data management, and cannot afford to develop a data curation infrastructure themselves. The participants raised another serious concern of who will take the responsibility of supporting the preservation of data in the longer term. The participants from the Centre of Ecology and Hydrology (CEH) are supported financially by the government to provide and manage data centres. This is due to the fact that the CEH data is of national interest and provides societal benefits to the public related to, for instance, land use, water, soil, and agriculture. In contrast, most of the other participants generate a lot of environmental data that can have a high impact on science and on communities but, due to cultural issues and the lack of funding for data management, those data become inaccessible to other researchers. In order to make this data accessible, there is a strong argument that the responsibility of data curation should be shared in trusted bodies such as universities, and institutional repositories. Hence, both universities and

institutional repositories collectively can play an important role to preserve environmental data for long term (re)use and research.

### 3.2.3 Overall Reflections

The overarching theme that emerges from this section is the obvious importance of data in modern environmental science. This breaks down into the following three key observations:

- Data is the lifeblood of contemporary environmental science and plays a key role not only in understanding the overall ecosystems but also helping resource managers (water, land, health and marine resources) and policy makers to shape their decisions and develop strategies about climate change respectively.
- The practices and technologies are clearly insufficient and suffer from either methodological limitations (old technologies) or technical and financial issues (particularly related to environmental sensors and IoT technology).
- There is an increasing variety of sources of data, which may lead to different levels of veracity around the resultant data.
- There is a lack of integrated solutions (e.g. distributed data repositories) for long term data preservation in environmental science; hence, data can lead to 'dark data' where it is not accessible or available to other researchers and hence is of low value.

## 3.3 Trends in Data Management: Openness, Collaboration and Integration

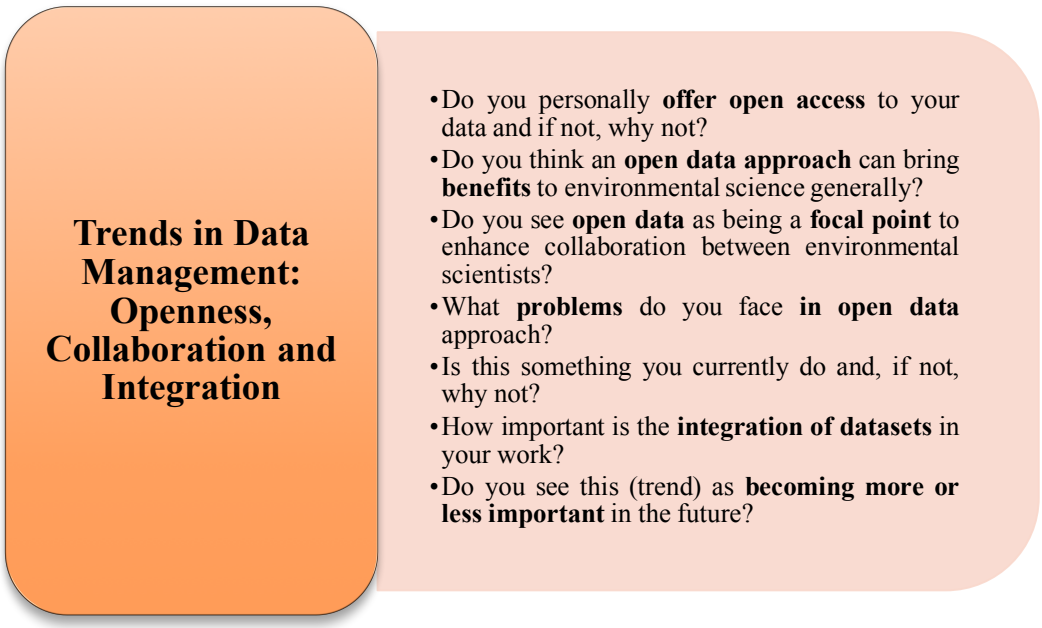


**Figure 3.4 Trends in Data Management**

### 3.3.1 Background

According to the Bromley Principles [179], *"full and open sharing of the full suite of global datasets for all global change researchers is a fundamental objective. Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers."* In order to exploit open data at its full potential it demands four essential requirements to be met. [180] describes these four requirements:

- Accessible - Data must be stored in such a way that it can be accessed quickly and without difficulty.
- Intelligible - The data must have a written description of the results of scientific work which should be comprehensible to those interested researchers who want to understand or possibly correct them.

- Assessable - The recipients of data should be able to assess the data, for instance, not only they are able to judge and scrutinise the nature of the claims the scientific work possesses on the basis of data but the competence and reliability of the claimant as well.
- Usable - Data should be available in a format which can be reused easily for many other functions. Contextual information such as metadata plays important role in usability of the data.

Open access to scientific knowledge has been practiced by many preprint servers, scientific journals, researchers' websites and worldwide institutional repositories and facilitated by the Science Commons for licencing. Funding bodies, policy makers including research councils, journal publishers, educators, and the general public are now pressurising researchers to adopt an open data approach [180-181].

In this section, we present findings on different trends in data management in environmental science, particularly focusing on openness, collaboration and data integration, as shown in Figure 3.4.

### 3.3.2 Main Findings

**In response to questioning about the culture and trend of open data**, the participant below explains how research councils persuade researchers to adopt an open data approach:

*"There is now a requirement that if you apply for funding to research councils in the UK, they want to know how about data will be managed not necessarily QC/QA on that data but how can it be accessed by other scientists or even general public which I think is right. The journals and research councils are pushing the environmental scientists hard into doing open data approach. Now most environmental scientists realise that they can't continue to get awards and grant money if they clearly not making their data freely available."*

Open science more generally makes the scientific information, methods and research results open and free to the interested researchers. Releasing scientific theories along with their experimental data to the public allows them to be strictly and thoroughly examined and corrected for errors if possible, making them refined or rejected [180-

181]. Thus, scientific knowledge progresses further through this open scrutiny and challenge.

The following participant recognises the role of open data in science scrutiny and progress:

*"I'd love to share my data with other interested researchers in my discipline. Because when you provide open access to your data along with your hypothesis and the procedure, the other researchers could validate, verify and sometimes rectify your hypothesis which is the best way of scrutinising and improving science."*

Sharing research data with others can enable the scientists: to reproduce or verify research, to ask new questions, to advance the state of research and innovation and to make available the results of publicly financed research to wider community [180].

The following participant explains the potential role of open data approach in endorsing the aforementioned rationale of data sharing:

*"I think there is an increasing expectation and of course demand of transparency in access to data, reusability of data and reproducibility and auditability of research results to underlying data. This is evidenced by science journals requiring DOIs to reference data behind submitted papers."*

All participants also agree with the fact that if a project is funded by the NERC or any other government funded body then the data collected in such projects should be freely available to the public. One of the participants identified this fact:

*"We absolutely agree with open data policy. We provide open access to almost all our data and encourage this approach. We accept it is a NERC policy which requires in principle that we've collected data at public expenses and it should be made open to all including the general public."*

Another participant identified the need of open data culture:

*"Yes, I'm a big supporter of open data approach and we want it desperately because the causes of environmental change and the response for the environment need a wide range of data to investigate. This can't be achieved within a close data culture."*

**In response to a question whether open data can bring benefits to environmental science**, one participant said, "*I think that's too obvious. The more you share your data the more you learn. When open data is advertised enough everybody knows about it then you can access data to help test a hypothesis or answer big questions. The more data you have the better ways you have manipulating those data the more insight you gain in science. So, I think there is no argument about that.*" The following participant emphasises how important is open data culture in order to avoid duplication of research:

*"Oh yes, it's long overdue when people just having to collect their own data every time they start on a research project was a silly idea when they were not only duplicating work it was done elsewhere in a lot of cases but also not being able to compare and contrast data or linking it up with other data. Well, you can see what happened but it was not a good way of doing things."*

Another participant pointed out the same benefit in terms of saving money on duplicating research:

*"Well, I say yes to open data because I'm concerned openness promotes better research, for scientists use each other data and know what other scientists or research groups doing at the same time. This can also save public money by avoiding duplicating experiments."*

Open data can also be useful in those situations where research is based on data that is not available or accessible yet. There are several reasons of this data unavailability, for instance, insufficient data, cost, scarce data collection instruments, and lack of advanced automated instruments or supercomputers etc. One of the participants, doing research on air quality based on ozone layer datasets, illustrates this case:

*"Yes, definitely we should have an open data culture in general and in environmental science in particular. One specific example I have in my research based on ozone datasets. There are several areas of the globe where there are no ozone data available. In some places, the scientists are not making measurements or can't do it for any reason. So, we don't have any information about the ozone air quality about those areas. But in other cases, such as China they are making measurements. Some of these datasets are made available online in real time so we could start grabbing air quality data from their web site today and undertake our research but we have no such information in the past years."*

**Answering the question of whether open data can be a focal point to enhance collaboration**, most of the participants were of the opinion that the more open you make data, the better it would bring together people to collaborate, as the following participant indicates:

*"The more scientists are prepared to be open about their data the more they are likely to collaborate. So, I think openness and collaboration are very much closely related, though I'm not sure which is cause and which is effect but yes open data should drive collaboration."*

The participant below explains how open data projects can potentially enhance collaboration among different environmental scientists:

*"I think there is increasing number of projects out there where people are realising that there exist some datasets that have not been fully utilised yet and so there is a value bringing back to order those datasets and extract meaning from them. Now that could be possible if scientists collaborate and look at these multiple datasets and do their own meta-analysis and then combine the results to get some interesting emergent results out of it and create a unique product or piece of research."*

Another participant pointed out when open data be a focal point to enhance collaboration:

*"I think it will be more like a paradigm shift that people need to embrace it. It will be a focal point when you are a part of a larger network (e.g. a EU grant) comprising 10-15 scientists who all produce different data and working hard to resolve data integration, compatibility and data quality issues around data. To collaborate in such environment to get the goals of the project open data could be a focal point."*

One of the participants identified the risks or fears involved in openness which could impede the process of collaboration among scientists but further concluded these risks are outweighed by the potential advantages:

*"There is a big risk and that is one of the fears why many scientists have been very reluctant in the open data game and that fear is it wouldn't result in collaboration, it would just result in anonymous scientists somewhere out in the world exploiting others'*

*data not giving them any credit or acknowledgment instead misusing their data. That is a threat which can't be ignored but I think the benefits of open data and the collaboration arises from it outweigh that."*

Only two participants disagreed with the idea that open data could enhance collaboration. One of them said, *"I don't think it will increase it any further. You collaborate with people because you know about their work; you may have built up a rapport with them. I don't think it will work as a focal point unless someone on the Web sees someone data that they never worked with before and say, 'oh instead of just taking your data I'd like to work with you' which is very rare."*

The other one rather made a contradictory statement that it will pose a threat to collaboration:

*"I think it might discourage collaboration if actually the data is just out there. If I could access the data freely and openly why would I need to go to the individual and collaborate with him or her? I could just exploit his/her datasets. So, I don't think it will initiate or enhance it. I think it would rather discourage it and will lead to less collaboration."*

**In spite of all these benefits, open data also poses some challenges and concerns**. The participant below raised the concern of requiring large data space and lack of computing skills to share data openly:

*"I would provide access to my data (processed files and scripts) if contacted, and I say this in publications. If I had a dedicated, large file space that could be accessed from the Internet, and some technical computing knowledge and skills of using application tools for providing access to data, I would have uploaded the data there. At least one of my PhD students is using GitHub for his code, and I would like to encourage this practice."*

Another participant expressed his concern that it would be an additional burden on him to spare time and make effort to distribute his data openly and freely with others while doing science and administrative duties at the same time:

*"I often don't provide open access to my data and the reason is because it takes a heck of a long time and effort to package data appropriately to be distributed. We already have awful administrative burden and other things we do then we can't do the science which is*

*why we do the job in the first place. So potentially if a job like data sharing or management takes a lot of time that would be an extra overhead on any sort of campaign."*

If data is not shared in a methodical and systematic way to extract meaning from it easily, it will be open and interpretable to only a small group of people having technical expertise [182] which could ultimately lead to poor data with no use or the issue of data quality. One participant pointed out this concern:

*"But it's not done in a systematic way. I don't follow a formulaic approach to say, 'right I've generated a data set, this is how it's going to be laid out and whatever data management program, these are quality control flags conforming to some national standards'. So, this is an issue come back to quality control."*

Sometimes data in one domain might be very helpful to one researcher but it might be very complex for researchers in a related or interconnected domain. In addition, if data is provided without procedural or contextual information it becomes very hard to understand. This concern was raised by the following participant:

*"I managed somehow using my social contacts to collect data from a researcher in a related sub-discipline in my research. But gosh, I can't understand the data, not at all. The data is provided as an Excel sheet having no descriptions of the results or procedure or any other contextual information. The data talks gibberish in my discipline and sounds like only numeric values which I can't understand what those numbers mean."*

Scientists live in a competitive world where they are working hard for personal promotion and incentives, financial benefits and winning research grants. They are governed by the number of publications and they do not necessarily want to give intellectual capital away, so this could constrain open access to their data until they publish their work. The following participant explains this:

*"There is always a competition for publications that you want to be the person whose name is against that piece of science. So, we don't share that particular data with anybody else until we publish it and then it's available unless we are working in a consortium where you need to pass the data around to get the final publication. So, we are very selfish on our data until we get our names against it."*

Cost is another issue when it comes to provide free and open data access. The participant below explains this:

*"We've never been compelled to have open access until now, that's only coming within certain funding streams of research. Everything is going to open up here and everything is the norm. So yes, we'll go toward open access but the barrier to that is cost. So, in some cases we go free to journals but it costs us a huge amount which is the main impediment in providing open access to our data."*

In addition to an embargo period mentioned earlier, which was the main concern of all participants, the participant below identified one of the privacy issues that can arise while releasing contextual information related to data that can deanonymise the location of the sample and status of individuals (for instance farmers or land owners) where it is collected:

*"There are few caveats to open data approach. One is usually the embargo period where it's one or two years to enable scientists to publish before their data becomes publicly available. Secondly, when it's not in public interest to do that. An example of that would be a countryside survey which is a 1km2 randomised survey across the UK and the dilemma there is in revealing the location of sample square. It would create a situation where a lot of other researchers would want to go to those squares and start taking additional information or measurements. Now that actually would be a breach of an agreement with land owners about releasing data. But more importantly it begins to give a biased sample square because a lot of people are working in the same square and that would have influence on what's recorded. So, it would no longer be a valid random sampling square. Also, it would piss the land owners off and they are more likely not to give us permission to go back to that square."*

Licensing becomes a tricky issue in provision of open data access when scientists are working in an environment where data belongs to other individuals, groups, consortiums or any third party. The participant below is willing to share his data but can't practice it because the data is not their own but belongs to third party having licencing restriction on them:

*"As a NERC Centre, some of our datasets are available under open government licence which means free to use just acknowledge. But a lot of our data is still under a separate licence. That because primarily we might use third party datasets to produce it, we might use our own survey data, we might use Met office data to produce datasets. So, licencing is a bit tricky issue because if we might have taken say Met office data and created some derived products from that then we will have to include the particular licencing setup so that is taken care of. So, we've to deal with a number of restrictions that may be necessary just because of the nature of the data."*

Some organisations sell their data for commercial interests or any other financial sustainability and the scientists have to pay for it to pursue their research. The following participant mentioned this:

*"One of the problems regarding open data access in this country is that some organisations like Met office would only release data as a cost or if you have research arrangement with them. So not all data in this country is open access, even some has been collected on public funds like the countryside survey data the CEH do is not generally available openly. I think now the older versions are available but not the current one."*

The above participant went on explaining the reason when he was asked why some organisations charge for their data sharing with others when it has been collected on public funds:

*"Because going back over successive governments some of these organisations would require funding their activities by selling their products, for example, the Met office will charge for weather forecast for specific purposes, it will also charge for his rainfall data even though a lot of data collected by volunteers would provide them without payment. They used to have processing cost so they are required to compensate some of their processing cost in order to make them financially less dependent on government funding."*

Environmental science is an integrative, collaborative and interdisciplinary field which comprises many other sub-disciplines. To perform collaborative and large-scale synthesis for answering complex environmental questions, it requires integration of data from different sources. The results of these synthetic analyses play a key role to inform decisions regarding sustainable management of the natural environment [139].

Therefore, it is important to effectively integrate different datasets from different environmental sub-domains to gain insights. However, such integration is challenging because it needs to understand differences in methodology, representation formats, and terminologies [144].

To discover the significance of data integration, **participants were asked about how important data integration is in their interconnected sub-disciplines**. All of them emphasised and recognised the value of data integration in their work, as explained by an environmental chemist below:

*"It's very important. Thinking about some of the international reports about the state of certain pollutants, integrating datasets from different regions around the planets is essential to make sense of the global atmospheric issues. We also really need a handle on one of the quality controls and ultimately quality assurance that the data you're viewing in one dataset is compatible with another dataset of the same pollutant but acquired by different labs in different location."*

A soil scientist mentioned how important data integration in his research is:

*"It's very important because we bring together a whole range of datasets, e.g. hydrological data, soil data, climate data, data on population changes, land used changes to come up with a single coherent model of the soil sub-surface and make predictions about the future."*

The participant below, working on the delivery of long-term, large scale monitoring and experiments for the collection, provision and modelling of biodiversity data, noted the significance of data integration in his research:

*"It's hugely important because we are being expected to address complex issues related to the social and economic drivers of change and also the consequences. It's not good to saying 'well, UK's system is changed in a certain way and we know what's causing it and that's very interesting, thank you very much. We've to answer the so what question and the so what question, involving linking different data up to other areas of science in terms of what's the downstream impact of the change on an upland peatland area and who's it affecting and even worse what's the economic benefit or disbenefit of what's going on, poses enormous challenges in terms of data collection in the blackout of valuation."*

A climate scientist mentioned the significance of data integration in his research:

*"Data is usually central to my work, and I need to be able to integrate climate model data, and satellite and ground based observations. Even if I am improving or developing a new model, I will need bringing different datasets together to test that."*

Some of the participants identified the reasons of complexity around integrating different datasets. One of them noted:

*"Integrating different datasets from various sources is a complex task for us because we use different methodologies, a variety of instruments, and record different types of observations. Converting such disparate data into a common data representation model and then understanding the meaning of those datasets is an arduous task."*

All participants need combining different datasets together from various sources to have a unified view in analysing a research question. However, they also raised concerns about the complex task of data integration. All participants noted that data integration requires a lot of technical skills, effort and time. The following participant identified these concerns:

*"The number and types of data we produce are vast. Our science strategy is based on multi-disciplinary research which requires bringing different datasets together but what we haven't conquered is how to bring interconnected areas of research together easily on the basis of those data. You can do it on a science project using some semantic web techniques such as linked data, vocabularies and ontologies but that requires considerable skills, time and cost which makes it hard to persuade funders to spend a lot of time (5-10 years) and money to see a fruitful result. If we have a short-term gain in one area of science and show them (the funders) the benefits of data integration the other science areas will say we don't want to do this, we want to do this. So, it's very difficult to persuade not only the funders but the research groups as well because we are vastly different having a large number of heterogeneous datasets."*

In summary, all participants realise the fact that they need to study the interconnected disciplines of environmental science in an integrative and collaborative manner. However, to do so, they recognise that they must share their data with others. To address complex questions of environmental science around data, they want to

integrate different datasets from various interconnected sub-disciplines to get a unified view. However, for most of these participants data integration from such disparate sources is an arduous and time-consuming process due to certain reasons. Some of these reasons include the differences in their methodologies for data collection, types of observations they record, conversion into a common data representation model, different metadata associated with data and finally understanding the semantics of data.

### 3.3.3 Overall Reflections

To summarise this section, all participants realise the need for a paradigm shift in environmental science towards open data. They noted that open data can strengthen the way science is done and scientific knowledge can be improved or (rejected) through scrutiny and critical analysis. Most of the participants were willing to embrace an open data culture and recognised the potential benefits it would bring to environmental science and in enhancing collaboration among scientists. However, they identified that this approach also raises some technical, sociological, financial and legal challenges and concerns they are confronted with. Some participants mentioned the issues including the lack of technical skills, time investment, efforts requirement, publication rights, data misinterpretation, receiving no proper credit, incentives or attribution, and cost; others were more concerned about legal issues such as licencing and privacy issues involved.

Regarding the open data paradigm, they recognise the need to change the way science is recognised and the whole cultural aspect of the organisations and institutions. It is true that scientists have some serious concerns about open science but it is equally true that they do not practice it just because of the lack of understanding and awareness about the benefits of doing it. In order to adopt an open culture and make it common practice, all participants recognise the need to motivate, educate and train all those communities having or generating data.

The important theme that emerges from this section is the shift towards open data culture in modern environmental science. This breaks down into the following three key observations:

- Environmental scientists realise both the need and importance of open data culture and get benefits from this, however this also raises some technical, sociological, financial and legal challenges and concerns. This is rooted in observations across all interviews.
- There is a trend from data silos and individual working practices towards more integrative, collaborative and open science, and to enable and realise such a shift is a hard but important challenge to address.
- The underlying process of integrating different datasets from various sources is an essential task in analysing the data to make sense of it; however, this poses technical challenges and hence requires computing and computational skills and inevitably extra training.

## 3.4 Interdependencies in the Long Tail of Environmental Science

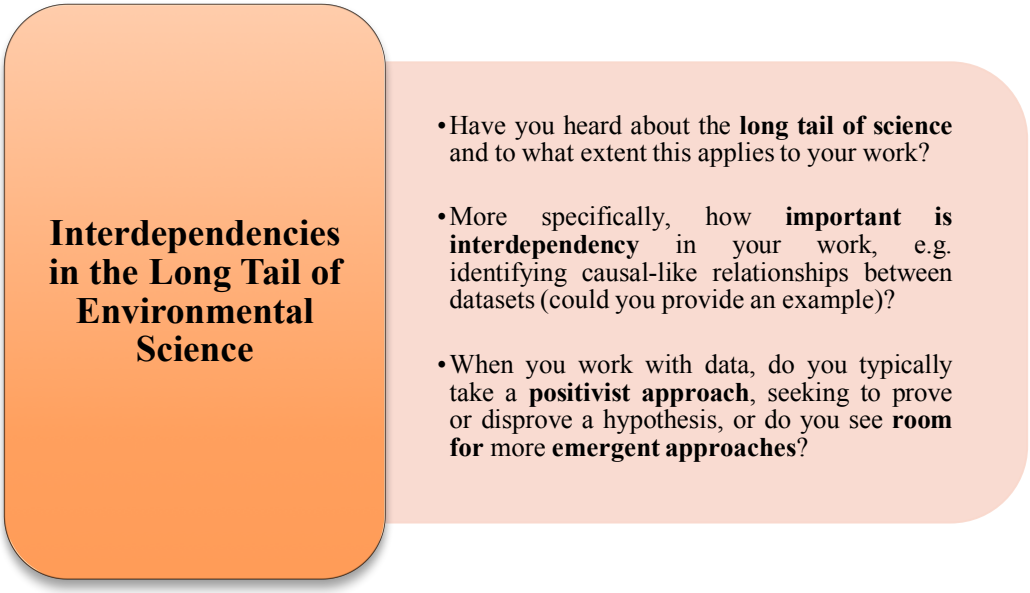


**Figure 3.5 Interdependencies in the Long Tail of Environmental Science**

### 3.4.1 Background

In environmental science, there is a need to investigate the key differences in environmental data where datasets tend to be smaller, more heterogeneous and where the main interest may rely on interdependencies between datasets representing

different environmental facets. This long tail data is characterised by wide diversity, hand sampling methods, non-uniform (unique) procedures, mostly individual curation and lacking community standards for data structures and metadata with no maintenance and seldom reuse [3]. One of the key challenges in the long tail of environmental science is to gain a better insight into interrelationships between different environmental facets and hence to understand the interdependencies in environmental ecosystems. Different industries, for instance, agriculture, water, tourism and urban development, among others, have conflicting demands which arise usually from the 'silo management' and consequently result in a development by one industry negatively impacting the other. Furthermore, due to the availability of advanced measurement technologies, a large amount of data has been generated. There is value in this data if it is examined in an innovative way. Hence, there is a need to seek more emergent approaches to deal with such complexities and interdependencies around data.

The aim of this section is to draw on three key aspects: (a) investigating where does the data in environmental science fall in the big-data-small-data spectrum (in other words, does environmental science fall in the long tail of science or is it in the transition phase towards big science because of automated instrumentations and increasingly large collaboration groups)? (b) discovering and understanding the associations between different environmental ecosystems and how they affect each other; (c) examining whether scientists still practice the scientific method of (dis) proving a hypothesis or seeking some emergent approaches to discover new facts and patterns in environmental data? These questions are summarised in Figure 3.5.

### 3.4.2 Main Findings

Most of the participants were not familiar with the phrase **'the long tail of science/data'** but when it was explained to them it became clear. When asked to what extent does it apply to their work, mostly said, *"yes, it does absolutely"*. As one of the participants, a geologist, explained his data:

*"I think it's probably very typical of the work that we do. There are probably some larger datasets that we also work with but the majority of data we have is fragmented, small and are spread across in various data types and formats."*

The following ecologist described his long tail data by pointing out one of the important characteristics of it:

*"The volume of data that we have is probably small but the variety of it might be quite big. We don't do terabytes of data; they mostly exist in climate science. So, the volume is not an issue but the real issue we are facing is the vast heterogeneity among datasets."*

Another ecologist, who was already aware of the long tail science, characterised his science:

*"Yes, I've heard about the long tail of science and most of our science we do is typified by all characteristics of small science which is usually short term, done by individual scientists or small research teams, diversified data, mostly hand data collections and personal computer data storage. So, the science we do is the real epitome of the long tail of science."*

An environmental chemist, similar to the above, has already got an idea of the long tail of science. As he explained:

*"I have heard of the long tail of science and it definitely applies to my work. I'm in that tail, i.e. I don't produce large datasets but very small one, mostly generated through traditional and manual sampling methods and unique procedures and consisting of many different kinds. These datasets often give the most interesting and valuable science regarding pollutants in the environment."*

The following soil scientist provided a very good explanation of both aspects of his science:

*"I haven't heard the long tail of science before but the way you described it finds my science mainly in the long tail. Though, I'm also part of a research project collecting global datasets on soil, I won't say the datasets are too large but it's collected through a mechanised way with uniform procedures and stored in internationally agreed formats and prescriptions. So, there is some element of the head (big science) but probably the majority of my science lies in the tail with small and heterogeneous data, collected usually through hand generated field sampling and processed through our own methods."*

A limnologist narrates her science in this way:

*"I think the long tail of science definitely applies to my work. We have a kind of big span of heterogeneous data where we have really some small datasets for very small projects having tens of data collection points. I'm also currently involved in a project where we've sensor networks which is generating huge amount (probably terabytes) of data because it's collected every four minutes. But this project is not part of any international collaboration or big project teams (like high energy physics or astronomy data) who have uniform procedures and central data curation. So, I'd say yes a lot of our work tends to fall towards this long tail."*

The volcanologist below explains the long tail of science in his research:

*"I generally work in the long tail. My datasets are usually small having different types because most of my work is field based, I'm restricted. I collect these datasets from various sources using different equipment. The problems I look at are relatively spatially small because I'm trying to understand small scale volcanic processes implying spatio-temporal changes which are geographically limited instead of global."*

A scientist in biodiversity relates his work completely with the long tail:

*"My work or research entirely fits in this long tail of science. The only new thing which is starting to move away toward slightly big science is our next big greenhouse gas project which is a multidisciplinary research involving several research teams on national scale. We'll generate a huge amount of data through a mechanised approach and uniform procedures and will curate it in a central data archive."*

A biogeochemist briefly summarises her science in big science- small science spectrum:

*"Most environmental science is in the tail and 'dark matter', a few exceptions such as climate science. I work mostly in the tail and trying to develop platforms to get stuff more in the tail. To be honest, environmental science is so fragmented, there are still a lot of individuals in different sub-disciplines who haven't had collaborated in big projects nor generated massive data, so, they are horribly down in the tail."*

To discover the interdependencies between disparate datasets, **we asked participants, "do they identify causal-like relationships between their datasets and how much it is important for them",** one of them, an ecologist, pointed out its significance:

*"It's an inefficient use of the data resource if we don't discover the causality among environmental variables and it's essential to understand these relations because the questions we've to ask are more and more complex and do have a necessity involved in linking up disparate areas looking at the interdependencies between them.*

Another ecologist describes the importance of interdependencies in his research:

*"It's absolutely essential to understand the drivers of change, the interactions between those drivers which are often very complex and the affects that have on ecosystems and the environment and the affects that have on the benefits to society."*

The above participant continued:

*"So, you can see there are a lot of interactions, a lot of interdependencies between data that need to be modelled and understood, for instance, the association between agricultural change and water quality. We broaden this out in terms of needing to understand how patterns of land use change in the landscape. There are many different stakeholders in the landscape be they foresters, tourists, nature conservationists, agriculture people, just businesses or people living there. So, you have this broad range of people who not only drive change in an area but also reaping benefits from it."*

The microbiologist below identifies the need of discovering interdependencies between disparate datasets in his research:

*"It's very important actually. The studies we do on catchments require data generated by molecular and microbiologists. We then relate that data with environmental data collected by the environmental agency, Met office, farmers and Defra. So, in our study we have got data about rainfall, river flow, river height, nutrients in water, number of animals in the catchment etc. We also move to do a kind of epidemiology as well where we are asking people about the diseases in the catchment and then looking for their potential causes and associations by analysing those data. So, getting all these different types of*

*datasets and joining up together to derive some interrelationships is very crucial in our study."*

The soil scientist below describes the relationship of soil with water and explains how much soil science affects hydrology:

*"What we've in water depends on the soil entirely and also things that go along with the water like sediments, nutrients and other pollutants depend completely on the interactions with the soil."*

Another soil scientist noted the interactions between land use, rainfall and water:

*"The interrelationship in our study is very important because we are working in all scales. So, in our national scale work, we are looking at changes in land use, driving micronutrient cycles, e.g. in our diffuse pollution work we are looking at changes in hydrology, rainfall events that drive sediments, nutrients etc. To interpret the nutrients data, we need the hydrology data, we need the rainfall data i.e. the intensity of rainfall and quantity of rain water. So, if there is a big rainfall event you'll get more water flowing through the channels, you'll get more sediments mobilised which end up with more phosphorous moving in rivers."*

The scientist below undertakes his research on the association between atmospheric science and soil science:

*"There is a big link between atmospheric science and soil science. For instance, we collect a lot of data on the pH of rainfall where we are interested in the measurement of sulphuric acid and nitric acid in the atmosphere. The acid gases e.g. sulphur dioxide and nitrogen oxide come from vehicles, fossil fuels and factories and are released in the atmosphere. When it rains these gases are deposited in rain and acidify the soil which leads to aluminium toxicity in soil which negatively affect plants. These waters come out of the soils and get into the rivers make the rivers very acidic and ultimately harm the aquatic life."*

The following climate scientist is looking for the interdependent relations between different facets of the climate:

*"This is very important e.g. in climate model evaluation. For instance, can we infer or even describe causal relationships between atmospheric composition change and climate change? If we can identify mechanisms that these links operate through, we can use this to see if the same mechanisms operate in the models and are the models right for the right reason. To find answers of these questions is very essential for us."*

A biogeochemist, looking at the interrelationships between land management, soil and water, identifies the significance of interdependencies in her work:

*"That's our job, that's what we do all the time. For instance, what is the change in land management that has degraded or improved soil quality and how that has benefited or changed water quality. So, these are three separate datasets that we are looking at to understand the interrelationships between them. This is what ecosystem science is, that is what biogeochemistry is, understanding that these things are all linked together and how they impact each other either positively or negatively."*

When participants were asked about their philosophy/methodology they take around environmental data and **whether they adopt hypothesis driven approach or looking for some emergent approaches**, most of them replied, *"both"*, as one of them mentioned:

*"Mostly we try to take a hypothesis driven approach. But to be honest, sometimes we just collect data and then will look at it to derive some interesting things. And some people argue that you'll not discover very much if you constrain yourself just to do hypothesis driven research that a lot of new discoveries are made by just measuring a lot of things. I agree with them, so, I'll say we do both."*

The following participant quite often practices hypothesis driven science but mentions the fact that hypothesis-driven science is not the only way to do science and there is always scope for emergent approaches:

*"I constantly do a positivist approach and a lot of my science is hypothesis driven that usually starts off with a sort of general question and then we shape it into a hypothesis and go for its testing. Though I'm a bit weak in emergent approaches but sometimes I follow a grounded theory where you build up data that tells you something without*

*necessarily challenging those data. So, I think there is room for emergent approaches but I'm very much rooted in driving hypotheses most of the times."*

One of the reasons shifting towards data-driven science is the availability of huge data. The participant below does hypothesis driven research mostly but in his new project having collected a lot of data he is seeking more emergent approaches to discover new facts:

*"The way I work is I setup an experiment with a specific experimental design that allows me to test a specific hypothesis with the statistical methods. So that's very much a formulated kind of way but now we've been creating so much data that we have had not in before. We have greater opportunity to mine data how can I call it exploratory kind of work looking at relationships and patterns in data so the things are moving forward for us from hypothesis driven approach to more emergent approaches."*

One of the participants more often starts off with a hypothesis but then further explores the data for some interesting patters:

*"Most of the times, we've a hypothesis we should do so we collect data to prove it using some statistical methods. However, there are cases where we've just curiosity so we just collect data and see what happens looking for some patterns and associations among different variables to see their response and then we create a story based on those analyses."*

The participant below is flexible and adopts the scientific method on the basis of question in hand:

*"I guess it depends on the data a bit really or what's the question you are looking for. Some things suit to hypothesis driven approach whereas others you can end up with doing more analysing data and looking for some interesting facts not discovered before and things you may find completely different that you didn't really realise that it's going to be there and that's very interesting that the things you really began with. I think it's good to be flexible and dynamic. So, I do both."*

Most of the scientists generally prove or disprove a hypothesis; however, because of the availability of more data, scientists are now looking for emergent approaches. One of the participants identified the need of emergent approaches in ecology:

*"Our science is based on hypothesis driven approach that still is an effective and easy way of doing things but it's quite limited when it comes to understanding complex ecological interactions. Now increasingly we have been looking at techniques which enable us to deal with that kind of environmental complexities. So, I think we are open to new approaches to do this. I was brought up with in terms of experimental design and sampling design which were helpful at times but really don't help solving some of the more complicated problems in environmental science. So, there is a need of looking for more emergent approaches."*

Because of using automated digital instruments, which collect large data over spatio-temporal scale, the following participant seeks more emergent approaches instead of doing hypothesis based science:

*"We spend our lives on fishing trips which you'd call it emergent approaches. So sometimes we go and test a hypothesis but now having large volumes of data using automated instrumentations we very often go and what's called fishing trips where we look for unexpected trends or relationships that will tell you something new that you didn't even think about before and I think that's perfectly acceptable way of doing science. You know it comes from unexpected surprises that people alert to and then go and explore it further. So, we do both but mostly emergent approaches."*

One of the participants truly supported emergent approaches:

*"We'll very much take emergent approaches or look to develop those kinds of approaches but I think culturally there are a lot of scientists who are quite suspicious of those approaches. You've to work hard to convince them of the need for things like data mining."*

The following participant noted the usefulness of emergent approaches:

*"Sometimes I work purely descriptively, for example, say the modelled and observed trend in surface ozone concentrations is X. However, emergent approaches are useful. Finding*

*new relationships and patterns between climate variables in a model (almost serendipitously), which then are also found in observations is a holy grail."*

### 3.4.3 Overall Reflections

In the big-science-long-tail-science data spectrum, most of the science done by these participants fits in the long tail of environmental science. The majority of the participants are doing science on a small scale spanning individual scientists to small research groups or small laboratories. Most of these scientists do not follow standard or uniform procedures. More often, they use local research methods varying from one sub-field to another with some adaptation. In terms of bigness, their datasets are small in most of the disciplines, except climate science which can produce relatively large datasets through simulations. In terms of diversity, the data is very heterogeneous with no widely accepted standard data format. Regarding the interdependencies, our findings also identified how important is to understand the complex interactions in the environmental ecosystems and some of the interdependencies which negatively impact each other, for instance, how intensification of agriculture can reduce water quality and ultimately affect aquatic life. Furthermore, it is true that most of the scientists still practise the hypothesis driven approaches. However, it is equally true that because of their curiosity and having a rich set of data from advanced measurement instrumentations they are now seeking room for more emergent approaches to find some hidden facts and significant patterns among the data. They are now looking for more collaboration with computer scientists and technologists to exploit data science techniques to get more insight into their environmental data.

The fundamental theme that emerges from this section is the clear significance of long tail data in environmental science. This breaks down into the following three key observations:

- The long tail of science is absolutely the core of environmental science and is coming up in all these interviews.
- To discover and understand interdependencies among disparate datasets representing different environmental facets is increasingly important in understanding overall ecosystems. Again, this is consistent across all interviews.

- Because of the advanced measurement instruments that generate more data, there is now a trend towards more data-driven science to look for interesting and emergent patterns among different datasets and turning them into knowledge.

## 3.5 Technology: Opportunities

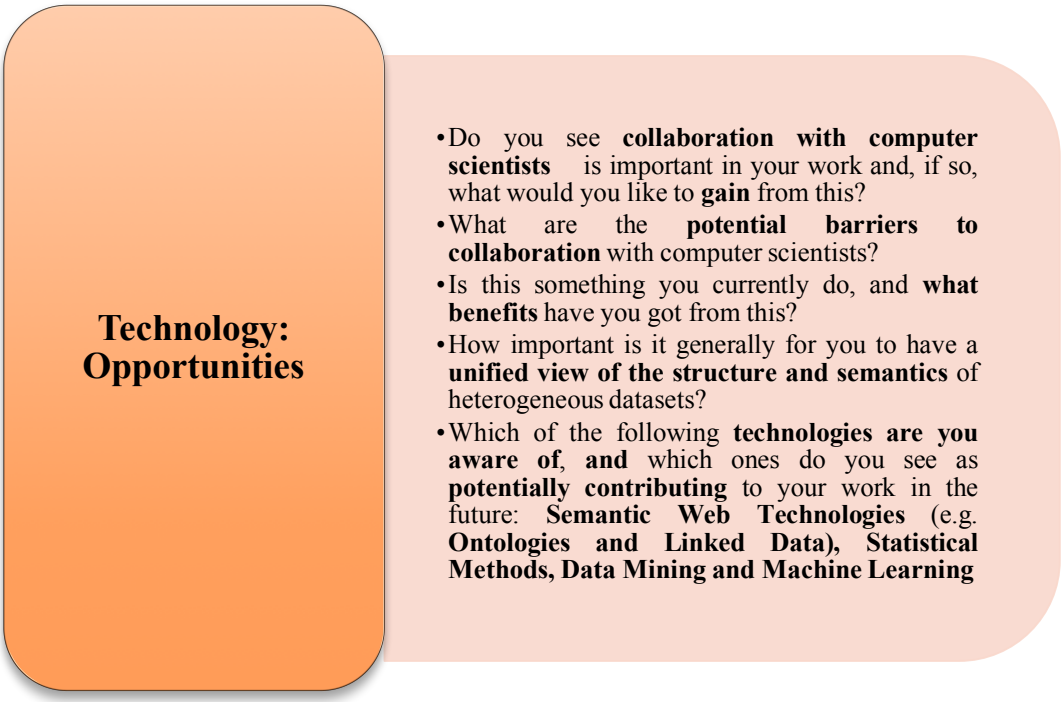


**Figure 3.6 Technological Opportunities**

### 3.5.1 Background

There is a real potential in interdisciplinary collaboration between environmental science and computer science owing to the issues of complex and heterogeneous nature of environmental data. This collaboration is very crucial because environmental scientists' knowledge of computational techniques lags behind the state-of-the-art in computer science. Hence, there is a need of bringing both communities together.

In this section, we present findings from the interviews, selecting responses to open-ended questions regarding collaboration between the two communities, benefits and potential barriers of this collaboration and the computing technologies the environmental scientists are benefiting from or aware of their use, as shown in Figure 3.6.

### 3.5.2 Main Findings

To investigate to what extent environmental scientists can exploit potential computational opportunities, **participants were asked about the importance of collaboration with computer scientists and the potential gains it can bring to environmental science**. All of them were in favour of working with computer scientists, as one of them stated:

*"Definitely, I see some value in computer science especially in understanding about ways in which different types of data might be available in my field and you might get access to those kinds of data."*

Usually, most of environmental scientists are not adept at computing skills to integrate different datasets and expedite their work. This fact was identified by one the participants:

*"Yes, it's very important because they have got skills we don't have and we are very unlikely to get those (skills) very quickly. I'd hope by collaboration we can start to create integrated systems that can speed up my work and my workforce and join together datasets in a way that is new and imaginative leading to new insights."*

Data management is one of the important aspects in data-driven environmental science. Because of the wide use of advanced instrumentations and IoT technology, environmental scientists want to collaborate with computer scientists to leverage these resources and manage their data effectively. As one of the participants mentioned this fact:

*"I need to work with computer scientists because they are the experts potentially in terms of data management, deploying environmental sensor networks, and how to program and the whole gamut of computer services or instruments and interactions with GPS etc."*

Efficient and new ways of data analysis is really important to environmental scientists. They have started looking for innovative techniques for dealing with the data to save their time, effort and money. As one of the participants explained its significance:

*"Absolutely it is critical. We might understand how environmental sensors and data loggers work but our experience and knowledge is rubbish in comparison with computer scientists. They can tell us new computing opportunities for data processing, analysis and research. They can also tell us what will work and trying to find things that won't work which will save not only time but also money. In addition, they are also better mathematically skilled than we are."*

Another participant, who has already worked with computer scientists and benefited from it, identified almost the same reason for collaboration:

*"We have spent some time with computer scientists working on one of our projects and definitely it brought some benefits to us. I guess they bring new ways of working with data and new tools to process those data we won't be aware of and that's useful because you can do things you hadn't thought you could do before. So, the biggest gain is just having somebody who's got computing skills to play with your data in an efficient and innovative way."*

It is not only data processing and analysis techniques but computational technologies such as cloud are also important to store environmental data and models. The following scientist, working in the data centre at the CEH, explains the importance of both computational and data processing techniques:

*"It's very important mainly because of our interest in developing capabilities with data centre. There are two aspects of that (a) there are still just basic computational approaches in use to data representations and (b) there has been the informatics side. So, overall computational technologies and techniques are emerging in terms of how to store and access information either semantically or through some sort of big data techniques."*

A climate scientist, who used HPC and other advance computational technologies to run their simulations, recognises the growing need of these technologies in their research and the dependence of their work on computer scientists:

*"Oh yes because the need for middleware in terms of accessing stuff across the cluster, cloud and high-performance computing systems for processing capabilities in our research has been increasingly growing. So, there is both that computational capability side and middleware side because we are not the people to write that sort of code so we are much more dependent on computer scientists."*

A key area of innovation known as data science is emerging in environmental science to achieve new scientific insights through a new integrative science. The participant below identified the need of data science in their discipline:

*"Yes, I think it's absolutely important because the computer scientists now are bringing a lot of techniques and tools that offer greater opportunities for exploiting the data and the environmental scientists may be aware of this but may not be. So, I think that environmental science is in need of a new kind of science which is to some extent the hybrid of traditional science and data science and I believe the scientists are now showing signs of trending towards this which is really good."*

**Answering the question regarding potential barriers to collaboration with computer scientists**, all participants identified technical jargon as a major hindrance to collaboration, as one of them mentioned:

*"There is always a problem in interdisciplinary project and that is understanding each other's vocabulary. You have to spend time to understand the technical jargon and the different platforms the scientists work with."*

Another participant, similar to the above, raised the same concerns in collaboration between the two communities:

*"Sometimes it requires both sides understanding new terminologies, the jargon that has become so abundant in both sciences requires a long time for your brain to remember all those terms. So, that's one of the biggest problems - this small technical jargon problem becomes the biggest, I think."*

Some of the participants identified a very important factor that environmental scientists are oblivious of the benefits of computer science in their discipline. In

addition to the language barrier, this fact was identified by a few scientists as another reason for the lack of collaboration between the two disciplines. As one of them said:

*"I think firstly it's the lack of knowledge what can be gained from computing sciences. So most environmental scientists have very poor understanding of what computer scientists can do for them in order to be benefited. Secondly, they are obviously two different disciplines and whenever you get that there is always a language barrier in terms of how they discuss particular issues."*

Another scientist, similar to the above one, termed it as 'ignorance' of the environmental scientists:

*"Ignorance is one of the main problems, we don't know what the computer scientists can do and what do they can offer and vice versa. So, it's all about the lack of communication between the two communities. Language barrier is the second main issue when you work with people from other disciplines."*

Most of the participants have just started collaboration with computer scientists; some will start in near future and a few have done in the past. All these participants have either benefited in the past or are likely to get benefits, as the participant below explained:

*"We've got a lot of benefits from this collaboration. I don't think we can do the projects having data without somebody technically very skilled having innovative ideas to analyse the data in a better way."*

The participant below, similar to the above, has benefited a lot from collaborative projects with computer scientists, as she explained:

*"Definitely we have benefited from collaboration mainly in computational approaches, data representation, semantics, and data science. We do collaborate with computer scientists to get new insights into computational issues and skills we don't have in environmental science."*

**When the participants were asked how important is for them to have a unified view of the structure and semantics of heterogeneous datasets,** most of them said, "extremely important", as one of them mentioned:

*"It is very important. A lot of work we are doing nowadays is about dealing with different formats of our heterogeneous datasets and the inconsistencies involved in it."*

Some scientists find the interpretation of their heterogeneous data really hard if there is no unified view of both the structure and semantics. A limnologist explained:

*"I think that's very important in terms of understanding what you are looking at when you've got the data. I mean a lot of the problems that I had recently with my datasets was just knowing how to interpret my data without having a unified view of both the structure and semantics."*

The majority of participants recognise the significance of a unified view of the structure and semantics of heterogeneous data but at the same time they also raise their concerns about the difficulty coming with it. Talking about the importance of a unified view of semantics and the difficulties involved in it, an ecologist mentioned her concern:

*"It would be lovely to get it but I know it's a nightmare. If you go and look into ecology and soil science, it's absolutely a nightmare. I agree it is very important but I also agree it is a lot of work to get it done in a complex area like ecology."*

Similar to the above, another scientist emphasised the importance of the unified view as well as raised the concern about this cumbersome task.

*"It is massively important to be able to bring different datasets together because the science that we are doing now to answer society's big challenging questions needs us to work in the interdisciplinary ways and make use of integrating different datasets in a unified way."*

He further continued and identified the difficulty:

*"The difficulty comes when you want to bring very different data streams together regarding one environmental problem in an effective and unified way to understand that data easily."*

One of the scientists raised another concern that getting a unified view of semantics of different datasets is a time-consuming job. He further said that however it is so

important to them that they cannot even work with their data unless a unified view is achieved:

*"Obviously it is very important but people working for me spend a lot of time to work out a unified view of the datasets. So, the work that we do is vital because you need to be able to work with multiple sets of data in a unified way and if you don't have that view of the structure and not getting the semantics involved in those disparate datasets you can't do anything with it."*

Some of the scientists are getting frustrated w.r.t. the semantics annotation of their data because they really need it but they cannot do it owing to the issues of lack of computing skills. Getting this unified view of semantics looks almost impossible for them, as one of the ecologists termed it a Babel fish:

*"Well, I would say it's very important but not practical in my discipline to expect that to happen very quickly. It would be like asking for the Babel fish. The Babel fish was something you put in your ear and it will give you an instantaneous translation of tons of languages spoken across the universe, yes, it is very important but I don't believe that it could happen in my science in near future."*

The participants were asked about **how much they are aware of different computing technologies/tools and techniques (dealing with the data) including ontologies, linked data, data mining, statistical methods and machine learning**. They were further probed which of these technologies have they used or are currently using and do they see any of these technologies potentially contributing to their work in near future. There was a mixed response. Five participants already knew all these technologies, as one them said:

*"I am aware of all of the technologies you mentioned."*

Regarding practicing and utilising the above technologies, statistics is the only tool that has been used by all participants, as one of them said:

*"We use loads of statistics, it is one of the main things for a lot of data analysis, and for most of environmental scientists, the only way to get publish all our data is analysed using these techniques."*

Some of the scientists were not aware of Semantic Web technologies particularly ontologies and Linked data. However, once the definitions of these technologies were explained and made their meaning clear to them, all of them said they definitely need these technologies:

*"I have not heard of these technologies but after your explanation of these terms I can see the value of its exploration in my area. There is a real potential of these technologies in my area of research."*

A soil scientist explained:

*"I have not heard of all these but these technologies, particularly ontologies and Linked data, sound really important and we need someone who could explore these technologies in our science. It could be someone from computer science like you who could work it out for us."*

He further continued:

*"I am sure there is a potential but I need someone who has got these skills to interlink soil, land use, hydrology and other datasets in my work. There are some attempts e.g. NERC is trying to link some environmental data together but whether they are using these technologies or not, I don't know."*

Half of the scientists are already aware of Semantic Web technologies including ontologies and Linked data but they have not used it yet in their areas. As one them mentioned:

*"Yes, I have heard of it. In fact, we have used ontologies in a collaborative project with bioinformatics. I think it is more widely used in life sciences but we have not used it yet in our research."*

One of the scientists, who already knew about ontologies, identified their significance in his area of science:

*"In my work, understanding semantics of heterogeneous data is extremely important. I need to know the meaning of the datasets and the relationships between them to make sense of it."*

Some of the scientists, in collaboration with computer scientists, are going to start using ontologies and Linked data in their new projects:

*"We are starting to use it in a project to have linking, trying to develop a platform for data for catchment management. We are trying to use or probably develop ontologies which would make our data kind of consistent. So, yes, we are starting to use it."*

Some scientists are desperate to develop ontologies in their area but due to the lack of computing skills and the heterogeneity of data they cannot do it themselves. They are just waiting for collaboration with computer scientists to accomplish this important task:

*"We are very much interested in developing ontologies but we can't develop it without the help of computer scientists because of the lack of computing skills and having so much disparate data."*

One of the scientists, an ecologist, identified a concern of not seeing any practical use of ontologies or linked data in his discipline:

*"I have heard a lot about ontologies and would be really happy to have one in my area. But I have a serious concern that people have invested time and effort and are doing something but have not yet seen any of much practical use. I have not seen it delivering anything of much practical use yet."*

He further explained and raised the same concern about linked data as well:

*"I would love to have inked data in my area and I would like to see any implementation of linked data but again so far no practical working example exists that can prove the concepts. It never seems to happen a real demonstration of linked data in ecology."*

When the participants were asked about the usage of data mining and machine learning, only a few of them know about these technologies:

*"Yes, I have seen potentially machine learning could be used. I have not really used it but people use it trying on various datasets e.g. seismology is one area where that sort of things has been approached e.g. using cluster analysis trying to look at signals and group*

*them into families which represent specific processes but it is not the sort of the things I do."*

A biogeochemist explained that she has not used data mining or machine learning in her research but her colleagues have been using it:

*"I haven't done it myself but quite a lot some of my colleagues have done it for, say, what is changing in the data, what are the interdependencies between datasets. They are getting data and turning it into some new knowledge."*

Similar responses were observed regarding data mining and machine learning techniques. There is a real potential for both these techniques but again less awareness and lack of knowledge came up across all interviews.

### 3.5.3 Overall Reflections

All of the participants interviewed showed their willingness and enthusiasm to collaborate with computer scientists. They see certain reasons for this collaboration: i) environmental scientists get benefits and potential gains both in terms of technological use and deployments, for instance, environmental sensor networks and IoT technology, and in terms of computing skills and advanced data analytics; ii) however, they face difficulties to resolve technical issues particularly around data and get frustrated because of their lack of computing knowledge to overcome these issues; iii) their computing skills and the knowledge they have lag behind the current state-of-the-art in computer science; iv) finally, there is an emerging trend in modern environmental science towards more data-driven science, which is rooted in observations across all interviews and scientists are now looking for new data analytical techniques to discover interesting and hidden patterns in their data and make sense of it.

The barriers of collaboration between computer science and environmental science are partly cultural, partly organisational. In addition, the lack of understanding and language barriers are another major hindrance to collaboration.

Most of the participants are aware of Semantic Web technologies and are interested in their implementation in environmental science. They see a real value and potential in these technologies. However, still there is a lack of understanding and experience.

The essential theme that emerges from this section is the obvious importance of collaboration between the two disciplines and the real potential and need of Semantic Web technologies in environmental science. This breaks down into the following three key observations:

- There is an opportunity in terms of collaboration between environmental science and computer science to understand this interdisciplinary, data-driven and integrative science. However, there is a challenge of breaking the cultural, organisational and technical jargon barriers.
- Getting a unified view of the structure and more importantly semantics of complex and heterogeneous environmental data is very important.
- There is a real potential of Semantic Web technologies to understand complex and heterogeneous data in environmental science.

## 3.6 Technology: Barriers

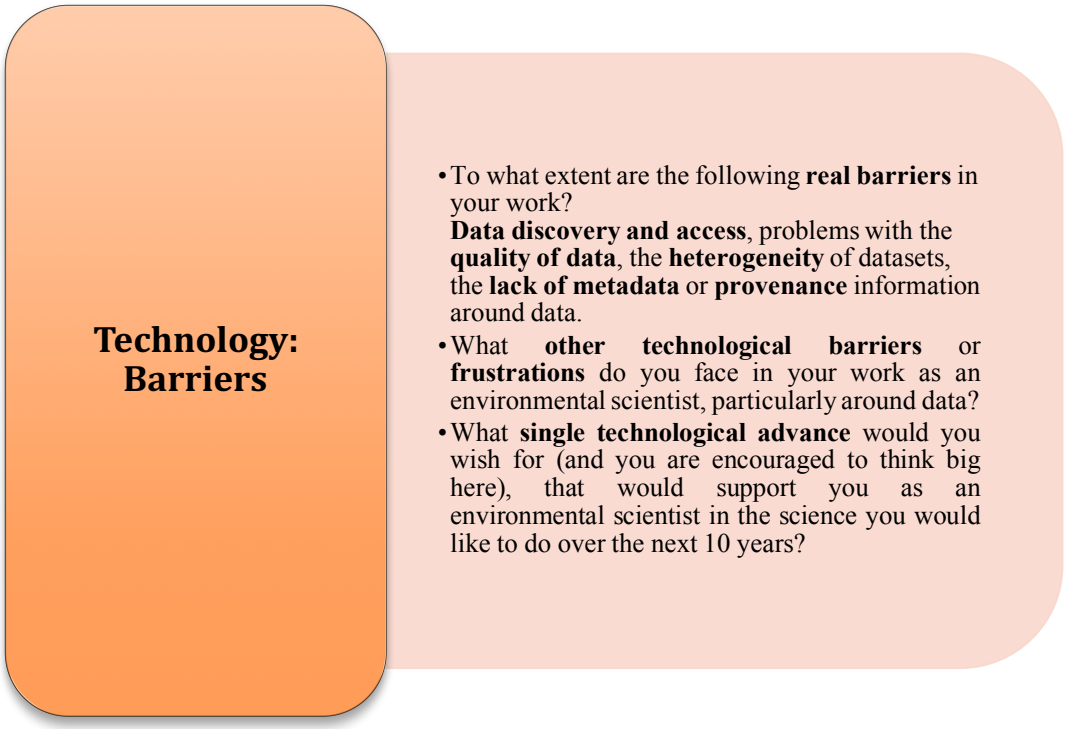


**Figure 3.7 Technological Barriers**

### 3.6.1 Background

Environmental science is an interdisciplinary field and comprises many other sub-disciplines. Because of its complexity, scale and heterogeneity, environmental scientists face some technological challenges including data discovery and access, data heterogeneity, data quality and data provenance, to name but a few. Data discovery enables us to locate the pertinent and available information in a particular knowledge domain and is one of the issues in science in general and environmental sciences in particular because of the vast scope and complexity of the discipline. Furthermore, making data available on the web does not mean easy discovery. One of the factors that causes data heterogeneity issues in environmental science, is the variety of interconnected sub-disciplines and the shift of contemporary research towards interdisciplinarity and collaborativeness. Data quality issues in environmental science arise when scientists get inaccurate or missing data in a dataset owing to the

use of different types of sources including malfunctioning instrumentation, inadequate documentation of data values and data entry errors. Data provenance, also known as lineage or pedigree, is described in databases as the description of the origins of data and the process by which it arrived at the database [183]. Data provenance in environmental science is of paramount importance, which enables researchers to determine the authenticity, quality and reproducibility of the data.

In this section, findings are presented regarding data challenges in environmental science including data discovery and access, data heterogeneity, data quality and provenance, any other technological issues or frustrations and any single technological advance environmental scientists would wish for to support their science, as shown in Figure 3.7.

### 3.6.2 Main Findings

**The participants were probed whether data discovery and access is an issue in environmental science**, most of them answered, *"yes, it is a barrier"*:

*"Yes, it is a problem. Data is very disparate; it is stored in different places so people don't know where it is. You learn over time and your career. It is getting better where people are trying making data available but it is still a barrier."*

A couple of scientists mentioned that it is not an issue in their research. One of them said they create their own data and their research is not dependent on data from others. The other one said it was an issue in his research ten years ago but it got better now due to their local data management centre. However, it could be an issue for wider data access:

*"No issue for me, but may be for others. Ten years ago, it was a big problem but now it is getting better because of our local data centre. It could be a problem for a wider data access."*

The other one said:

*"It is not an issue for me; we generate a lot of our own data so we don't wait for other people's data."*

If data is available on the web, it does not make it easily accessible or discoverable. It still requires a lot of time, efforts and energy to find the desired data one is looking for. This concern was raised by one of the participants:

*"Today everyone is saying data is available on the Internet. Well, it might be true that most of the data is there but putting simply data on the Internet does not mean it is discovered easily. Mostly I find it so taxing and laborious. For instance, sometimes when I look for particular specie on the web, I find hundreds of results, which is so hard to find data about my own. There should be an efficient and automated way to find the data of your interest quickly and easily."*

A couple of participants reported that they can access only that data which is published or uploaded on the web and freely available:

*"Yes, it is an issue. We have access to data that is either published or went up on the internet but still not all data is out there, I mean not accessible. We are limited with the amount of data all the time, we have access to what is there and we don't have access to what is not there."*

**When participants were asked about data quality to know whether it is a real issue**, everyone responded, *"yes, it is a major issue"*, particularly in case of (re)using other people's data because they do not know a lot of things regarding this data, e.g. what instrument was used, who collected the data, what QA/QC mechanism was used, what was the confidence interval, etc. Environmental scientists collect data in different environmental conditions using a variety of instruments, methods and sources. Due to these factors data quality issue occurs:

*"It is really a big issue, particularly reusing other people data because you don't always get all the information you need to understand, e.g. how the data was collected, what methods have been used, so those kinds of things, so it hard to know how things are comparable."*

Some of the participants mentioned technical malfunction as a very important reason of getting incorrect data which further leads to data quality issues, as one of them said:

*"Some sensor networks create bad data at certain times, how do you know that the sensors are working correctly and how they were calibrated. Such kind of issues lead to data quality issues."*

One of the participants identified that environmental conditions is one of the main factors that results in data quality issues:

*"Quality of data is always an issue and entirely varies. I work outside if the environmental conditions are good then you have great data and you don't have to worry about it, you have to check the quality once you have got back. Others day you go out and the conditions are dreadful and the data quality is not good because the instruments do not work correctly. It is not just good and bad instruments; it is also environmental conditions which can affect the quality."*

Environmental scientists usually contact the person who generated the data regarding data quality issues. However, the problem gets worse when scientists do not know the originator of the data and the anomalies in data are hard to correct them without the originator's help:

*"Normally we analyse the data and find the problem and contact the person directly via email and just have a discussion how the data was collected. Sometimes you don't find out the originator of the data and spend years trying to work out which datasets are good and which are bad."*

**Data heterogeneity is another major issue that has been reported by all participants**. One of the reasons for arising this issue is using different data formats *and models:*

*"Yes, it is a serious issue specially if there are different frequencies of data. That is a real hassle, when you have got an important dataset having different formats and data models, it is tricky for those who don't understand the best way to deal with it programmatically, then it a real issue. And it would not be an issue for me, it would be an issue for all environmental scientists."*

Some of the participants said data heterogeneity issue arises because of using different data capturing instruments and methods. Maintaining data having different data formats is hard for them:

*"It is coming our way and yes, it is a big issue. We have to take all forms of data and store it in a much common way as we can. We can't keep disparate data formats because it makes it harder to maintain. What worries us is the various data capturing instruments and methods and I think that may lead to big issues in terms of heterogeneity."*

Sometimes environmental scientists use different terminology for the same physical quality, for instance, one scientist would use the term, say, nitrate, other would use $No_3$. Contrary to this, some scientists would use the same term for different concepts e.g. using the term temperature for air temperature and soil temperature. One of the participants identified this issue of semantics while integrating different datasets together:

*"It is a huge issue when you want to integrate different datasets and want to know the meaning of different terms or concepts involved in while bringing those datasets together because scientists sometime use different terms for the same concept or same term for a different concept."*

Data heterogeneity is a big issue technologically for environmental scientists, particularly when they want to bring different datasets together, because they need to have computing skills to address this issue, which they usually do not have:

*"It is a big issue technologically because bringing together very diverse datasets is a skilled task and I think environmental scientists don't have those skills typically to do that."*

The same participant continued:

*"So, heterogeneity brings two problems: either descriptions of what the heterogeneity is, or secondly, the technical skills required to actually integrating those datasets when they are typically in a range of different data formats and most scientists don't have the skills to bring those datasets together."*

A few participants identified the time and effort factors involved, in addition to computing skills, to address this challenge:

*"Yes, it is an issue and it takes a lot of time and effort to handle with such heterogeneous datasets. There have been some attempts in soil and land use to produce some standard formats for recording data but a lot of data has some bespoke type problems because of the lack of computing skills."*

Another participant noted that he writes his own code in Matlab to resolve this issue; however, it takes a long time and requires proper computing skills because the solution he has is not sufficient to handle all types of data:

*"To resolve heterogeneity among different datasets, I write code in Matlab to process different datasets because there isn't any standard software or tool. So, this is the sort of things that I would find working with computer scientists very useful because it takes me quite a long time, and obviously the code is still not sufficient to deal with all types of data."*

Some participants identified the fact that there is lack of techniques to address this 'nightmare' in any effective way. It stops them from doing their science:

*"Well, it is a nightmare that prevents you from even going there if you want to deal with. It is rather like the problems we talked about before we don't know any really good example so far where really heterogeneous datasets have been brought together in any effective way. So, it is such a problem that I am not really sure we would be able to tackle it at the moment."*

**The interview continued with questions around the lack of metadata or provenance information**. Most of the participants said that lack of metadata results in both provenance and quality issues. There is no standard way to produce reliable metadata and it would take a long time to have one such standard:

*"The data provenance is one of the first fields in your metadata because it ultimately affects the data quality and also if it is good the provenance would also take the standard that has used to produce it. In environmental science, we are a long way from having a reliable or standard way to do it."*

Most of the participants noted that when they draw data from other people or literature, they face a lot of lack of metadata issues.

*"Yes, it is a big issue, as I mentioned earlier there are very often many datasets from external sources which often have no metadata et all or not sufficient, so you won't be able to properly interpret it."*

Some participants identified a very important fact that if they get metadata that comes with a published paper, even that does not fulfil the criteria and hence is not at the level they need:

*"Everyone says it is in the paper and you go and look at the published paper but it is not nearly the level that you actually need. It doesn't serve our purpose."*

Only two participants mentioned that they do not have such issues because they either use the national datasets in their work, which are well-documented, or generate their own data:

*"We have been using national datasets in our current work. So, we are not really having any data provenance issue because they tend to be well documented data. Also, we have generating our own data so we haven't been relying on data from others."*

Another participant said he does not have any issue because the metadata is available:

*"It is not usually a problem in the kinds of things that we do because mostly people have recorded reasonably metadata, I think. In our area, more often people will know the people who have collected the data and metadata and will often go back to them and will ask them follow up questions."*

The provenance issue gets worse when the data is very old and does not come with sufficient metadata. It takes a long time to amend or recreate such data:

*"These days we are trying capturing metadata that we got 50 years of data which does not have the right metadata behind it and it is very time consuming to recreate that metadata."*

One of the participants noted a very important point:

*"Yes, it is an issue, but we are publishing guidance to researchers nowadays, these are the minimum info that you have to provide. If we had stuck with a really strict criterion we wanted, it would get better."*

**When participants were asked a question about other technological barriers or frustrations around data**, most of them said they have already described the issues they are facing. However, there are still a few issues some of the participants had not mentioned yet:

*"A lot of the sensors are not reliable, in terms of telemetry the signal quality is appalling, you can't actually implement it in the real world, particularly in harsh conditions. There are challenges all the way along that loop."*

A few participants, especially from climate science and soil science, mentioned their real frustration regarding the lack of high performance computational facilities such as cloud technology, cluster or HPC, to run their climate and soil models respectively. They noted that processing large datasets on desktops computers takes a long time to get done. Lack of efficient algorithms is another barrier in their science:

*"Processing large data is the biggest challenge we are facing. We need powerful computers and efficient algorithms to deal with it. Normal desktops are not efficient enough. We try to predict things that happen over large areas e.g. to calculate water flow in a big catchment area, we can't run that on a standard desktop computer, it might take a long time, and that is our huge frustration."*

To some participants, understanding interesting patterns among datasets and then turning them into knowledge in an innovative way is really important. One of the participants mentioned her real challenge about the significance of looking for new patterns and making sense from different datasets:

*"My frustration is, I don't have computing skills to deal with data analysis on different sets of data. I want to understand interesting patterns between different but related datasets, what is it showing, and converting it in a form to make sense from it, is a huge challenge for me in my research."*

One of the ecologists identified his frustration of adopting an old way of doing science:

*"So, my frustration is we are often enforced to adopt the old system for doing science, which does not give us time to step back and say, well look it is stupid if we kind of design the system properly we wouldn't have to keep doing this, we would be able to join up our datasets properly, we wouldn't be constrained by the current approaches which is archaic really. So, that is the frustration if I have time to step back and look at it that we don't move forward quickly enough on this to provide those usable methods that can be used by people who are not computer scientists and they are not necessarily technically adept but they do know what to do with the data whey they get it."*

One of the participants wants to maintain a sustainable backup of his data. He has lost his valuable data in the past. His real frustration is maintaining all different copies of his data trackable. He further says the technology is there but he does not have funds to do that:

*"Yes, maintaining a sensible and sustainable back up and maintaining all different copies of the data trackable is my real frustration because I do my best but I am not formally trained in that. I guess there might be some software or techniques for people like myself and we don't need necessarily the rigidity of a formal database and then getting data in and out needs to be quick and easy."*

One of the participants working at the data centre identified a very important barrier regarding the mismatch between data compliance requirements and the way scientists want the data:

*"We have to have structured things we are legally obliged to deliver the data in this form. The scientists don't like this, they want the data in the form they want to use it and it is not necessarily always in that form. This is where we are moving away from compliance to satisfy environmental scientists as much as possible."*

**The last question of the interview is about a single technological advance environmental scientists would wish for to support them in their science over the next ten years.** We collected one of most important findings, based on the answers to

this question, around data, techniques and the associated technologies to deal with the data. One of the participants said:

*"I'm not sure it should be a new advance, it would be more applying what is already known in environmental science and it would be creating a middle layer between myself such that when data came in, it could be analysed and integrated in a more efficient way. We need software that could automatically integrate datasets and if required interlinking it with other data so that scientists could focus on science rather than data manipulation analysis."*

Another scientist said almost the same thing:

*"We need a smart kind of database that could integrate and interlink our datasets in an easy and automated way, I don't know whether there exists such a database."*

Some scientist wished for software that could intelligently find some new patterns and derive new knowledge from the existing data captured from the sensors in the environment:

*"You cannot beat the impact that you have when you are trying to illustrate and see patterns, new knowledge and relationships between different datasets; and that would be really healthy in my work."*

Another participant wished for better environmental sensors, in addition to intelligent decision support system:

*"I want better environmental sensors that could collect data about the natural environment in real time, and then an intelligent decision support system to use that data and is able to reason over the data to inform some decisions at the end of the day."*

A soil scientist, working in collaboration with hydrologists, mentioned the significance of geospatial reasoning in their collaborative project. She said she wants to have software that could reason over spatio-temporal data to discover and understand spatio-temporal trends in the environment in order to be able to respond to these events in time:

*"My single technological wish is about having a smart system to perform geospatial reasoning about different events occurring in the natural environment regarding weather monitoring, land usage, geographic events, hydrology and soil science and pollution monitoring. We are always interested to discover and understand the spatio-temporal trends in the environment in order to be able to respond to the emerging trends or geographic events."*

One of the biogeochemist mentioned that they want a technology and intelligent software to perform geospatial reasoning for finding the answers of complex questions:

*"We need an intelligent system and a smart knowledgebase to find the answers of various kinds of complex queries anytime we want to retrieve, for instance, when is the right time to apply the fertiliser at the right place, when is the right time to sow the seed, have sheep been in the field and for how long, what was the soil moisture value during the intensive rainfall or flood and how long the flood did last for, what is the status of soil, I mean has it saturated or hydrophobic, and where are the high risk pollutants' zones etc."*

One of the participants wished to have all the data in his science at one place:

*"Ideally I need an access to all the data in my science and all the data at one place and that would really help me to focus more on my research."*

Another scientist wished for a technology to perform data integration and spatio-temporal reasoning of different datasets:

*"I want automatic integration of heterogeneous datasets and then reasoning over those data spatially and temporally. Imagine you have an intelligent software that is capable of say, you press a button and the datasets will be integrated with other datasets and also capable of spatio-temporal reasoning."*

When data is published using open W3C standards such as RDF and SPARQL, and can be linked to other people's data to discover more interlinked information is called five star linked data. One of the participants wished for five star linked data in his area of science:

*"I would like to have a five star linked data in my research and discipline and is a way forward."*

A volcanologist wished for a technology to automate different procedures and link disparate datasets for easy access and backup:

*"I would love the automation of procedures and linking my disparate datasets for easy access and backup mechanisms, e.g. where are my images from volcanoes from 1999 and it says you have 500 images and they are here."*

### 3.6.3 Overall Reflections

In this section, we report findings based on responses to the interview questions about different technological challenges and frustrations around environmental data. Firstly, most of the participants have real data discovery and access issues. Participants raised some technical, financial and cultural concerns regarding discovery and accessibility to data. A couple of participants mentioned that it is not an issue for them because their research is mostly based on their own generated data. Tackling data heterogeneity and quality of data are two core challenges reported by all participants. The main reason of their frustration to resolve data heterogeneity issue and achieve interoperability across datasets is the lack of computing skills. This has come out from observations across all interviews. The issues around data quality are partly technical and partly cultural. The cultural issues occur mostly because of following bad practices and getting no incentives or attribution for authoring well-documented metadata. Most of the scientists' metadata is not well documented. Lack of standards for data quality control and assurance is another reason in the long tail of environmental science. The real frustration of environmental scientists in their work, particularly around data, is dealing with integrating heterogeneous and complex datasets because they cannot focus on their research and most of their time is wasted to work out technical issues around their integration. Some of the participants are really curious to understand the spatio-temporal dimension of the natural phenomena. Spatio-temporal reasoning across disparate datasets is a real challenge and substantially important to understand the emerging trends in the environment to be able to respond to those potential events.

The overarching theme that emerges from this section is the obvious importance of data challenges in environmental science. This breaks down into the following three key observations:

- Variety, veracity, discovery and interlinking of environmental data are crucial and central challenges in the long tail of environmental science.
- The real frustration of environmental scientists is the lack of computing knowledge and skills to deal with complex and heterogeneous data in environmental science particularly w.r.t. integration. This is rooted in observations across all interviews.
- Reasoning about geographic events across space and time to discover and understand the spatio-temporal trends in the natural environment is another significant and fundamental challenge.

## 3.7 Overall Discussion

From the qualitative data analysis reported in this chapter, some important findings have been identified. First of all, there has been an emerging understanding of the role and potential of Semantic Web technologies in underpinning environmental science. From the study, it can be seen that Semantic Web Technologies can potentially play a key role in understanding complex and heterogeneous environmental data. Semantic Web technologies including ontologies and linked data can be used to describe these complex concepts and the relationship between them. Furthermore, as described in Chapter 2, Semantic Web technologies have the potential to reason over different data to deduce new knowledge, hence making sense of the data. This fact was also mentioned by one of the scientists working at the data centre of the CEH, "*We have a huge interest in semantics approaches and new techniques for processing disparate datasets to look for some interesting patterns or infer new facts from the data.*" Both ontologies and linked data can also potentially be used to integrate disparate datasets and to interlink the datasets with other external data sources, hence making an integrative, linked and open environmental data science.

Secondly, there is an obvious lack of understanding and experience of Semantic Web technologies in environmental science. Furthermore, there is insufficient awareness about these technologies, partially because of the lack of communication and contact

with computer scientists. Other disciplines including Life Sciences and most notably Bioinformatics have benefited more from such collaboration. The good thing that can be observed is the emerging trend toward more interdisciplinary collaboration between the two disciplines. The need for this came across in all interviews, as one of the scientists noted, *"the key frustration of almost all environmental scientists is the lack of technological skill set in environmental science addressing some often quite complex environmental challenges around data."* Hence, further research is required to investigate systematically Semantic Web technologies to understand environmental science around data in all its complexity.

Thirdly, there are some unique characteristics of environmental data that need to be considered in a solution based on Semantic Web technologies:

**1. Interdependencies between Disparate Datasets**

There often exist causal-like relations between disparate datasets representing different real-world phenomena and how one phenomenon can negatively impact the other. For instance, how the intensification of chemical fertilisation and/or the movement of livestock into lowland areas, combined with high rainfall and spring tides, can cause a significant transfer of nutrients and faecal bacteria into coaster waters and ultimately affecting water quality and aquatic life. Therefore, to exploit environmental measurement data at its full potential, there is a need to convert these low-level descriptions about the real-world phenomena into meaningful knowledge to get an insight into those events about the physical world.

**2. Geospatial Data Integration and Reasoning**

Geospatial data plays a key role in understanding our natural environment and is critical in application areas such as weather monitoring, land usage, understanding geographic events, hydrology and soil science and pollution monitoring to name but a few. The environmental scientists always want to discover and understand the spatio-temporal trends in the environment in order to be able to respond to the emerging trends or geographic events in a timely manner. The geospatial observations collected from sensors, if integrated and processed intelligently, can help in informing the decision making about the natural hazards.

### 3. Interoperability

Environmental science is a multi-disciplinary science, which comprises several interconnected sub-disciplines including ecology, hydrology, soil science, biogeochemistry, climatology, meteorology, oceanography and geography. There is a potential shift in this discipline where individual research scientists, working in silos, have been transformed into more integrative, interdisciplinary and collaborative research groups. In such environment, environmental scientists connected to these related subfields, work on a complex environmental problem and need to access and use data. They use their own terminologies, different measurement units, different data models and experimental designs, which leads to data heterogeneity issue. Data is obtained from diverse sources such as individual scientists, research groups, sensor networks, observatories and experimentations. Data might be stored in a structured form such as database tables, semi-structured such as XML and unstructured such as plain text, blogs and images. These scientists need to combine and understand datasets from connected fields in order to have a uniform view of the structure and semantic of heterogeneous datasets. There could be many approaches to resolve data heterogeneity issues. One approach is using the Semantic Web technologies that has the potential to help in addressing interoperability problems.

### 4. Data Discovery and Access

Data discovery enables scientists to locate the pertinent and available information in a particular knowledge domain. The issue of data discovery and access arises in environmental sciences because of the vast scope and complexity of the discipline. Data is available in a number of forms such as biological, physical and/or chemical; captured from observational, experimental and field data measurements; stored in different places such as Internet databases, CD-ROMs, institutional records, journal articles, national museums, public archives. Majority of the valuable data have no web connection and hence is unavailable to the broader community because of the ownership of data by individual scientists, national or international funded projects and public or government institutions. On the contrary, making data available on the Internet does not mean easy discovery, for example, looking for a particular data might bring tons of results in which the desired data is hardly found. Data access is restricted owing to the issues of geographically scattered environmental data, the

temporally sparse data, restricted access to numerical models, institutional hindrance to data access, for instance, compatibility issues, and financial hurdles such as paying huge amount to access the data.

**5. Data Quality and Provenance**

Data quality is another major issue that arises because of many factors including faulty instruments, naïve data collectors, bad environmental conditions, bad practices and lack of standards. Scientists do not follow a standard method of documenting metadata. They do manipulation of data and then it is not cross checked. Usually, data is not accompanied by rich and well-documented metadata. The issue exacerbates when the originator of data is not known. Finally, there is a lack of standard quality assurance and quality control methods that can absolutely prevent the introduction of errors and possibly correct the anomalies in data with minimal human involvement in the loop. A related issue that arise because of the lack of metadata is data provenance that serves as a foundation for data quality.

## 3.8 Conclusion

This chapter has examined the unique characteristics of environmental science around data through semi-structured in-depth interviews. The overarching themes that emerged from this study are:

- Data is the 'lifeblood' of modern environmental science.
- There is a potential shift in environmental science from 'data silos' toward more integrative and open data science.
- The long tail of science is a key characteristic of data related to environmental science.
- Collaboration between environmental science and computer science is important in order to overcome the technological barriers identified in the study above.
- Semantic Web technologies have the potential to understand complex and heterogeneous data in environmental science.
- Data heterogeneity, geospatial reasoning, interdependency between disparate but related datasets, discovery and access, and data quality and provenance are the key

data challenges in environmental science that must be addressed in any data management approaches going forward.

The work in this thesis focuses in particular on Semantic Web approaches specifically for streaming data targeting the data needs of the Environmental Internet of Things project. The work particularly addresses the first three of the five challenges (Section 3.7), electing to leave data discovery and access, and data quality and provenance as future work given the size and complexity of these topics. The next chapter introduces a systematic way of exploring Semantic Web technologies in terms of building an ontological model as a basis for addressing these three key research challenges around environmental data.

# 4 Ontology Design

To address the three key research challenges derived from the qualitative analysis of in-depth interviews in the previous chapter, an ontology has been developed. This chapter provides an overall design of the ontology, which involves integrating and extending existing standard ontologies to form an overall framework.

More specifically, the goal of this chapter is to develop an ontological framework to describe the data stemming from the Environmental IoT Infrastructure (target domain), described in Chapter 1. The proposed ontology conceptualises various concepts and characteristics of the target domain and the relationships between them. The ontology is used to enable low-level sensor descriptions to be used as semantically enriched datasets. These sensor measurements have been captured from the sensor nodes deployed in the Conwy catchment, North Wales. Through this, the near real-time sensor data will be semantically enriched using the vocabulary of the ontology.

A collaborative and incremental approach has been used. The approach is collaborative because, during the ontology design process, the input of environmental scientists was used extensively. Initial domain knowledge was acquired from semi-structured in-depth interviews. In addition, several meetings were held with environmental scientists regarding the ontology design. It is incremental because an initial version of ontology was developed from the domain knowledge, acquired in the previous phase. The ontology was evaluated with real-time use-cases and refined. To

conceptualise related additional characteristics (such as thematic, temporal, and spatial) of environmental data, not covered by the initial version of ontology, it was further modified by adding new concepts and evaluated. This process was repeated until an improved ontology was achieved. The ontology is developed in OWL 2 (Web Ontology Language) using the ontology editor Protégé 5.2.

The rest of the chapter is structured as follows. Section 4.1 describes the goals of the ontology. Section 4.2 discussed the design criteria that have been adopted in the ontology. The proposed modular design of the ontology has been described in section 4.3. Section 4.4 discusses different dimensions of the ontology in terms of representation (thematic, spatial and temporal) of metadata in this work. The core modules of the ontology are described in section 4.5. Section 4.6 provides a summary of the chapter. Finally, section 4.7 concludes the chapter.

## 4.1 Goals of the Ontology

To perform a systematic investigation of Semantic Web technologies, an ontology is developed. The main goal of the ontology is to represent different concepts and characteristics of the target domain, and the relationships between them to get semantically enriched sensor measurements. This goal can be further divided into the following more specific objectives.

- To describe the thematic, spatial and temporal concepts of the data stemming from the Environmental IoT infrastructure in order to discover possible interrelationships and higher-level insights between disparate but related datasets.
- To resolve data heterogeneity issue by defining unambiguous ontological terms of data and their meanings and relationships in order to achieve semantic interoperability between different terms.
- To integrate and reason over different sensor measurements regarding different environmental facets enabling scientists to answer complex questions in environmental science.

## 4.2 Design Criteria of the Ontology

In this section, we briefly discuss the design criteria that have been adopted by the ontology. Four main design decisions have been made, which are described below.

### Reusing Existing Standards

As described in Chapter 2, one of the main reasons of ontology development is to reuse existing standards if available in a knowledge domain instead of building an ontology from scratch, and hence reducing the engineering cost and enhancing the potential for uptake. In this ontology framework, several existing standard ontologies including the SSN, Geo, GeoSPARQL, and Time have been reused and extended.

### Modularisation and Extensibility

In order to keep the ontology extensible and possibly to offer a high-level structure, a modular approach has been used where the ontology at lower layer uses/inherits the ontology at the upper layer, as shown in Figure 4.1. The main feature of modular approach is decomposing the process of building an ontology into more manageable components, hence enabling easy import of other ontologies in the existing model. Extensibility is another key feature of the ontology so that new concepts and modules can be easily added or removed while reducing time and effort. For instance, if the ontology requires the description of provenance information, the PROV-O ontology can easily be imported in the ontology. Similarly, the measurement unit ontology can easily be removed from the ontology framework and a new lightweight ontology of metric units can be built.

### Expressiveness, Reasoning Support and Performance

Expressiveness is one of the most important features of ontology design and development. The more expressive an ontology is the more reasoning support it provides. However, there is a direct relationship between expressiveness and computational performance. Increasing expressiveness will directly affect computational complexity, and hence will lead to inefficiencies. As mentioned above, OWL 2 (Web Ontology Language) [184] has been adopted, which is designed and standardised by the W3C. OWL 2 provides strong expressive power in comparison to other Semantic Web languages [185]. In order to keep a good balance between expressiveness, reasoning support and performance, the OWL 2-DL sublanguage has been used because it provides both sufficient expressiveness and reasoning support while preserving good performance. OWL 2-DL retains computational completeness

(all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time) [186].

### Lightweight Ontologies

Another design criteria regarding the ontology design is the size of the ontology, i.e. how big and complex the ontology is. Keeping minimal and lightweight ontologies in the IoT domain is very important for the efficient management of heterogeneous data and devices. In practice, we could have a large, complex ontology that might lead to computational problems, for example, reasoning. Nevertheless, making ontological commitment is more important [62], i.e. the ontology should be based on the consistent use of vocabulary to achieve consensus across communities. Hence, the purpose of the ontology in this work is to aim for more lightweight but extensible model that communities can agree with and which can be extended over time as concepts are deemed missing.

## 4.3 Ontology for the Environmental IoT Data

One of the main reasons of developing ontologies is the reuse of knowledge. When ontology is built in a specific domain, others can reuse it in the same domain for their own purpose and application. As said above, a modular design is the best approach where the ontologies are layered according to their scope. The proposed ontology in this work has adopted the generic ontology model introduced by Guarino [21], which provides a top-down approach for developing ontologies according to the level of ontological generality. Guarino's model is based on modular design that provides an easy integration of different ontologies making it suitable to be adopted in this work. The target ontology is an integrated model that is comprised of an upper ontology, a domain Ontology, a method/task ontology and an application ontology, as shown in Figure 4.1.

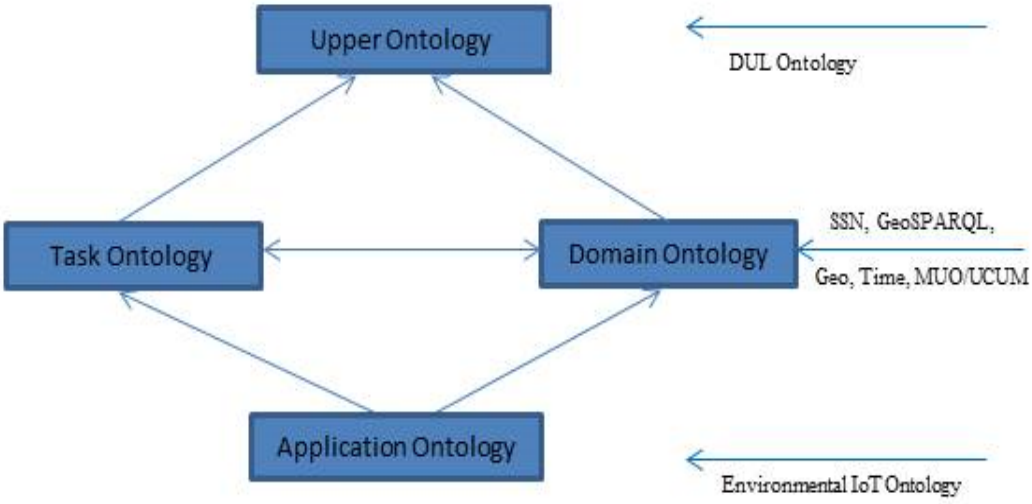


Figure 4.1: Environmental IoT Ontology Framework

The upper ontology, also called the generic ontology, captures knowledge that can be used across multiple domains. The main purpose of the upper ontology in the ontology framework is to provide wider semantic interoperability among domain specific ontologies [187]. The upper ontology, adopted and extended in the ontology, is DUL (DOLCE Ultralite) [188] that stems from the alignment of the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [189] and the Descriptions and Situations (DnS) [190] ontology. DUL describes concepts like object, event, process, situation, region, and quality, to name but a few. DUL is a lightweight version of DOLCE and DnS ontologies, which provides a set of upper level concepts that can serve as the basis for easier interoperability among many middle and lower level ontologies.

The domain ontology is developed for representing knowledge in a particular area of interest or domain (for example, environmental science, the IoT domain, bioinformatics etc.). In the ontology framework, several domain ontologies have been reused and extended to describe: information in the IoT domain (e.g. sensors, devices, observations, and feature of interests), information in the time domain (e.g. instant, interval, and duration), information in the space domain (feature, geometry), and information related to the metric unit system. The domain ontologies that have been reused and extended in the ontology framework include the Semantic Sensor Network

(SSN) ontology, the GeoSPARQL ontology, the Geo ontology, the Time ontology and the MUO/UCUM ontology.

The method ontology describes how domain knowledge can be used to perform specific tasks (e.g. diagnosis or scheduling). That is why this ontology is also called the task ontology. No method/task ontology has been imported in the ontology framework because the method/task ontology usually focuses on the problem-solving domains to accomplish a particular goal, for instance, expert systems.

The application ontology is designed to represent knowledge in a specific application and this usually contains both the domain ontology and methods from the method ontology [61]. The application ontology has been developed for streaming data stemming from the Environmental IoT Infrastructure. This IoT infrastructure targeted specifically local and regional environmental applications using inexpensive off-the-shelf technologies to understand the functioning of natural systems based on a network of sensors deployed widely across the landscape.

In the above modular design the upper level ontology can be reused across diverse applications because the more general an ontology is the more chances of reusability. The lower level ontology imports the upper ontology to extend knowledge and further enhance reuse.

## 4.4 Dimensions of the Ontology

In this section, different dimensions of the ontology for the Environmental IoT architecture are described. These dimensions describe different representations of knowledge of the target domain, for instance, where a particular event occurred, in which geographical location and at what time it occurred [191]. These dimensions include thematic, spatial and temporal. Consider an event: the sheep have been found in the field during Storm Desmond. The thematic dimension in this event describes what did occur (the sheep have been found), the spatial dimension describes where did the event occur (in the field), and the temporal dimension describes what time did the event occur (during Storm Desmond).

One of the core ontologies that is imported in the ontology is SSN. The SSN ontology is reused and further extended with additional classes, properties and relationships to

represent the information about the deployed sensor network such as sensors and their measurement capabilities, properties and feature of interests, observations and deployment and provenance of the sensors. To capture spatial and temporal characteristics and metric units of the measurements, the GeoSPARQL, Time and MUO/UCUM ontologies are extended respectively. The description of all these ontologies is given below.

### 4.4.1 The W3C Semantic Sensor Network (SSN) Ontology

In order to describe sensors and observations comprehensively, the W3C Semantic Sensor Network Incubator Group (SSN-XG) developed the SSN ontology in OWL2 [121-122]. The resultant ontology has 41 concepts and 39 object properties. The ontology is aligned with the DUL ontology, a lightweight ontology for modelling physical or social contexts. The SSN ontology inherits 11 concepts and 14 object properties from the DUL ontology for the alignment purpose. The conceptual modules, key concepts and relations of the SSN ontology are shown in Figure 4.2.

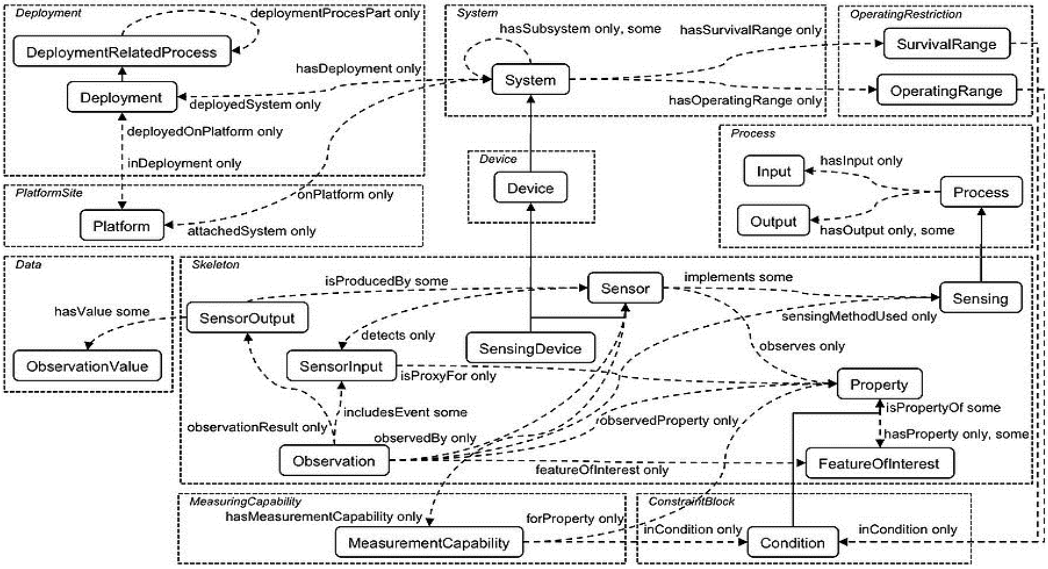


Figure 4.2: The SSN Ontology Conceptual Modules, Concepts and Relations [121-122]. The dashed rectangular boxes indicate modules, solid rectangular boxes indicate classes/concepts, solid lines (linking a class to another class) represent rdfs:subClassOf relations and dashed lines represent properties.

As can be seen from Figure 4.2, the SSN ontology is based around four main perspectives that are briefly described below.

- Sensor Perspective: where a sensor is characterised with a stimulus, sensing method, observation and capabilities, for instance, what is sensed, how it is sensed and what senses.
- Observation Perspective: where the main focus is on the observation that connects the incoming stimuli, the sensor and the sensor output.
- System Perspective: focuses on the system of sensors and their deployment.
- Feature and Property Perspective: where focus is on the sensed properties or the observations that have been made about them.

The SSN Ontology is based on the SSO (Stimulus-Sensor-Observation) ontology design pattern [123], which follows the principle of minimal ontological commitment [62] that means that the ontology should make as few claims as possible about the domain being modelled. This allows the ontology stake holders to specialise and instantiate the ontology as required, enabling reusability in a range of applications. The SSO pattern represents the relationship between sensors, stimuli and observations, as shown in Figure 4.3. Stimuli are changes or states detected by sensors in the environment. Sensors are physical objects used to perform observations by transforming incoming stimuli into digital representations. Observation serves as the nexus between the stimuli, the sensor, and the output of the sensor.

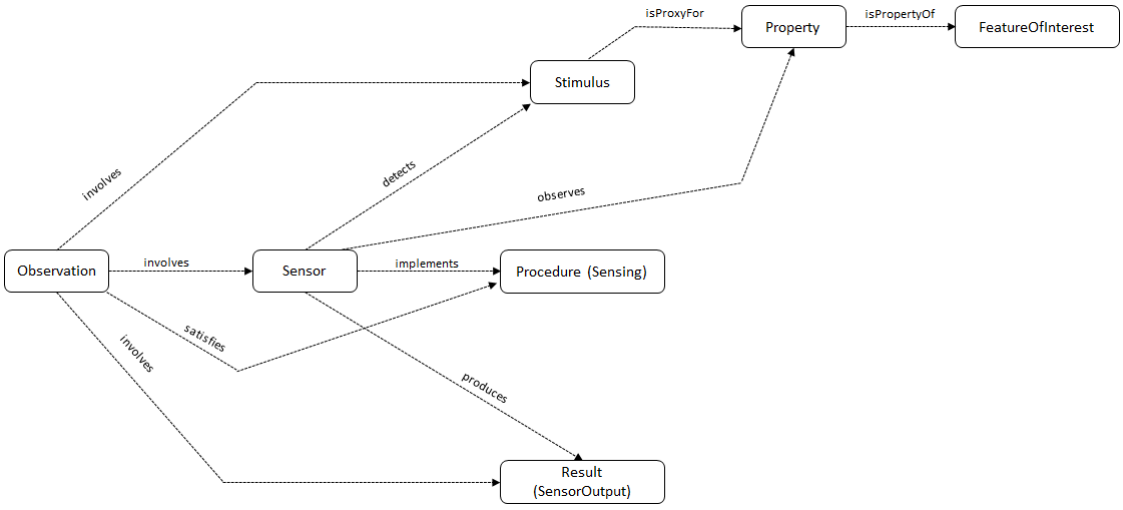


Figure 4.3: The Stimulus-Sensor-Observation Ontology Design Pattern. The solid rectangular boxes indicate classes/concepts and dashed lines represent properties.

### 4.4.2 Representation of Environmental IoT Metadata

The ontology developed in this work describes different features of data/metadata in terms of theme, space, time, and metric units. These representations are described below.

**(a) Thematic Metadata Representation**

The thematic metadata represents the main concepts or entities in a domain of interest. In our research, thematic metadata have been created mostly by sensors, for instance, about soil moisture, soil temperature, air humidity, air temperature, rainfall, and sheep etc., as shown in Figure 4.4.

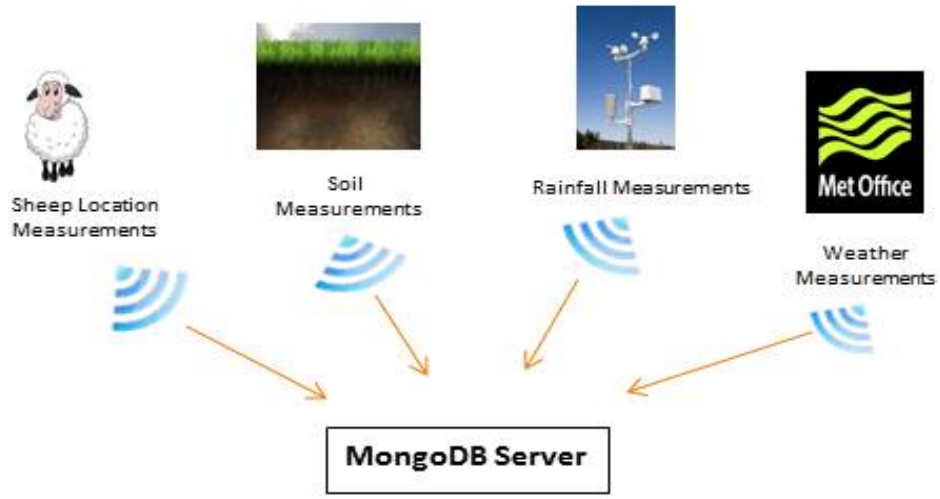


Figure 4.4: Sensor Measurements Stemmed from the Environmental IoT Project

**(b) Spatial Metadata Representation**

Spatial data plays a key role in our research to represent and analyse the geospatial dimensions of the environmental variables, for instance, where are sensors deployed, what is the location of sheep, what is the location of soil sensing node, what types of sensors are there in the river bank, what is the value of soil moisture at hilltop etc.

To capture geospatial coordinates and features, several ontologies exist to model spatial characteristics of sensor data. The ontology reuses and extends the WGS84, also called the Basic Geo ontology [192], because it is a standard lightweight

ontology defining a minimal set of vocabulary to represent the latitude, longitude and altitude of the GPS system. It has only one class called 'Point' whose instances can be described using the properties 'lat', 'long', and 'alt'. The benefit of Geo ontology, as said above, is its lightweight nature and simplicity, however this ontology cannot capture complex geospatial features such as polygon, rectangle etc. In order to overcome this limitation, the GeoSPARQL ontology [193], an Open Geospatial Consortium (OGC) standard, has been reused and extended in the ontology. The GeoSPARQL ontology describes information about spatial features and geometries and their relationships. In addition, GeoSPARQL provides some SPARQL querying functions and predicates for spatial reasoning [194]. This further extends the basic Geo ontology and provides different types of geometrical features, for instance, point, polygon, rectangle, triangle, line etc. These geometrical features use an object property including 'hasGeometry' and two literals including GML (Geography Markup Language) and WKT (Well Known Text), as shown in Figure 4.5.

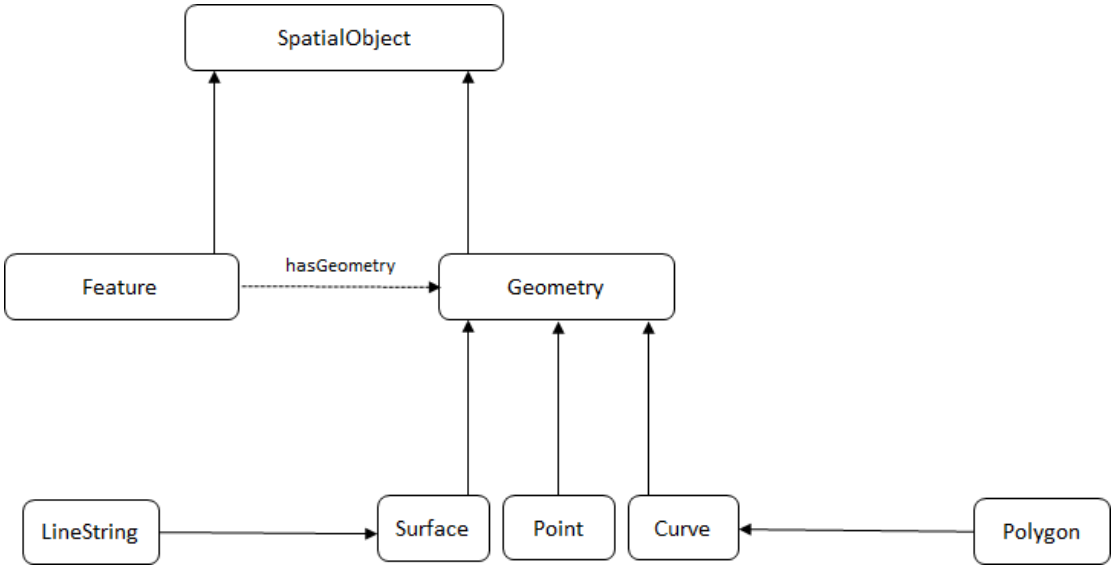


Figure 4.5: Geosparql Ontology. The solid rectangular boxes indicate classes/concepts, solid lines (linking a class to another class) represent rdfs:subClassOf relations and dashed lines represent properties.

In order to capture the geospatial features of the target catchment area, the area is divided into three different zones on the basis of geometry of the catchment. These three zones are named as Hilltop, Swale, and Riverbank. Three sensor nodes, namely A7, A8 and A9, are deployed in the Hilltop area. The Swale zone is instrumented with

seven sensor nodes, namely A3, A6, AB, AC, AD, AE, and AF. Sensor nodes A0, A1, A2, A5 and AA are deployed in the Riverbank zone. These three areas along with their deployed sensor nodes are shown in Figure 4.6.

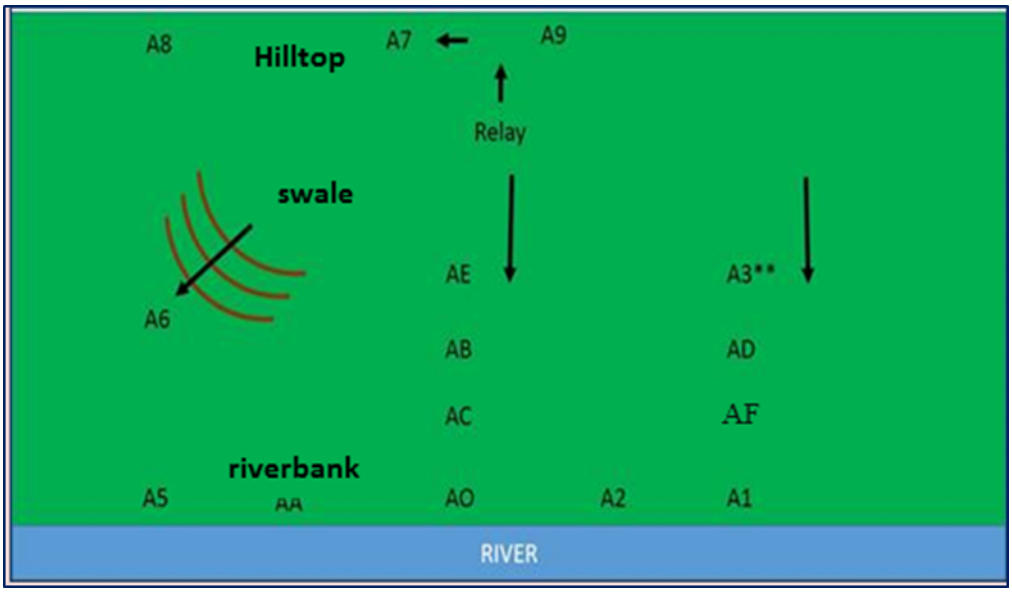


Figure 4.6: Sketch Map of the Sensor Nodes Deployed in the Catchment

**(c) Temporal Metadata Representation**

Temporal characteristics of sensor data represent knowledge about time zones and measurement timestamps. These attributes are as important as spatial in this research describing the information about the Environmental IoT infrastructure, for instance, when have the sheep been in the field, when did the flood or intensive rainfall event occur, what was the duration of the flood, when was the soil saturated etc. To address such queries about real world phenomena, the ontology reuses and extends the OWL-Time [195] ontology because of its lightweight nature and standardisation by the W3C. The Time ontology provides vocabularies to describe the temporal properties and relationships. The ontology also expresses the facts about the time interval and duration along with the datetime information. The Time ontology has two main classes called Instance and Interval with some additional properties like time: year, time: month and time: hour etc. The two main classes are shown in Figure 4.7.

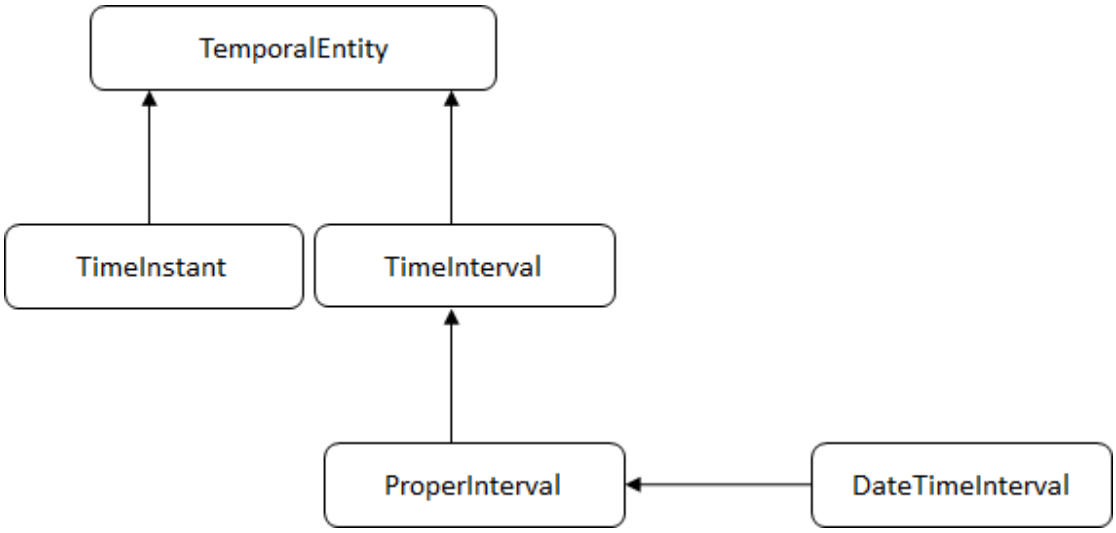


Figure 4.7: Main Classes in Time Ontology. The solid rectangular boxes indicate classes/concepts and solid lines (linking a class to another class) represent rdfs:subClassOf relations.

**(d) Metric Units Representation**

Quantitative measurements are incomplete if they are not presented alongside their associated metric units. Metric units are basic scientific tools to provide meaning to these measurements. Unit ontologies are also used to perform semantic interoperability and help in data integration. Currently, several unit ontologies exist and there is no consensus on a standard ontology. In the ontology, the MUO (Measurement Unit Ontology) and UCUM (Unified Code for Unit of Measure) ontologies are reused and extended to represent measurement units for physical qualities such as soil temperature, soil moisture, air humidity, acceleration, and rainfall etc.

**Namespaces of Existing Ontologies Used in the Ontology**

The ontology namespaces used in the ontology framework are listed in Table 4.1.

| Prefix | Description | Namespace |
|---|---|---|
| SSN | The SSN Ontology | http://purl.oclc.org/NET/ssnx/ssn |
| DUL | DOLCE+DnS Ultralite Ontology | http://www.loa-cnr.it/ontologies/DUL.owl# |

| | | |
|---|---|---|
| Geo | Geographical Location (Basic Geo) Ontology | http://www.w3.org/2003/01/geo/wgs84_pos# |
| GeoSPARQL | The OGC Geospatial Ontology | http://www.opengis.net/ont/geosparql# |
| Time | The W3C Time ontology | http://www.w3.org/TR/owl-time/ |
| MUO/UCUM | Metric Units Measurement Ontology | http://purl.oclc.org/NET/muo/ucum/ |
| enviot | Environmental IoT Project Ontology | http://www.environmental-iot.com/enviot_ontology/IotSemanticModel# |

Table 4.1: Ontology Namespaces Used in the Ontology Framework

## 4.5 Design of Core Modules of the Ontology

This section describes the design and development of core modules of the ontology, which extend the imported ontologies including SSN, GeoSPARQL, Time and MUO/UCUM. An ontology module is a small and interlinked conceptual fragment (component) of the ontology that can be considered as a self-contained and reusable component of the ontology preserving relationships to other ontology modules [196]. The ontology has been edited in Protégé 5.2 version. The core modules of the ontology are discussed below.

### 4.5.1 The Sensor Module

To represent different sensors of the Environmental IoT Infrastructure, the ssn:Sensor class of the SSN ontology is extended the to capture the descriptions of three main categories of sensors. These three categories of sensors are modelled as the subclasses of the ssn:Sensor class including AcclimarSensor, CampbellSensor and GroveSensor.

These three sensors along with the description of one particular type of GroveSensor sensor, i.e. enviot:GroveSoilMoistureSensor,(highlighted) are shown in Figure 4.8.

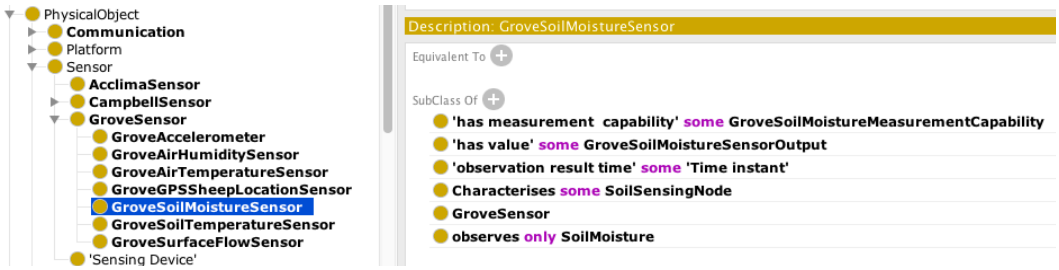


Figure 4.8: Description of the GroveSoilMoistureSensor Class

The ssn:Sensor class has two object properties: ssn:hasMeasurementCapability that describes the measurement capabilities of a sensor, which are expressed as an instance of a class, and ssn:observes that describes what property a sensor observes, for instance, soil moisture, air humidity, air temperature etc. In order to describe the measurement capabilities of a particular sensor, an instance of the class ssn:MeasurementCapability is defined, e.g. enviot:GroveSoilMoistureMeasurementCapability. To link this instance to its measurement capabilities, the property ssn:hasMeasurementCapability is used by creating an assertion on a particular sensor, e.g. enviot:GroveSoilMoistureSensor. As an example, the measurement capabilities of the enviot:GroveSoilMoistureSensor class are shown in Figure 4.9.

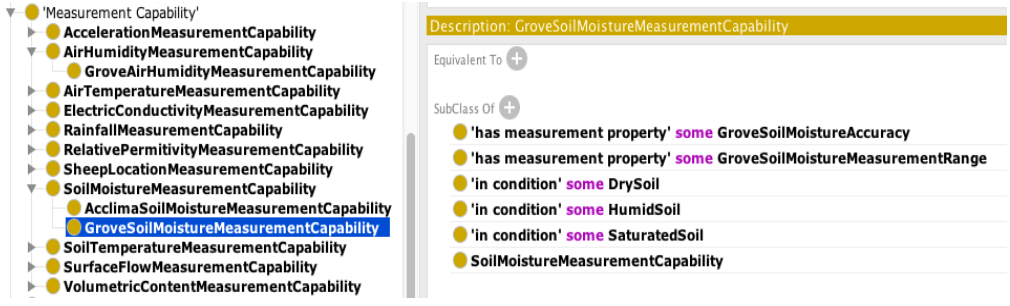


Figure 4.9: Description of enviot:GroveSoilMoistureMeasurementCapability Class

### 4.5.2 The Observation Module

Observation (ssn:Observation) is a situation (DUL:Situation) , which is produced by a sensor using some sensing method. Obsevation describes both an observed property or a feature of interest and a value attributed to that property by a particular sensor.

Hence, the result of observation is the output of a sensor. Several properties for the instace of the ssn:Observation class are defined, some of them are summarised below.

- ssn:featureOfInterest: points to the observed feature of interest, which can be any observed real-world phenomenon, for instance, soil, saturated soil, weather etc.
- ssn:observedProperty: points to any property observed by a paritcular sensor, e.g. soil moisture, soil temperature, air humidity etc.
- ssn:observeBy: points to a paritcular sensor that observed the observation, e.g. enviot:GroveSoilMoistureSensor.
- ssn:observationResult: points to the result of an observation, which is the output of a sensor, e.g. enviot:GroveSoilMoistureSensorOutput.
- ssn:observationResultTime: points to the time the result of observation became available at.

Extending the ssn:Observation class, a sub-class, called enviot:GroveSoilMoistureObservation is defined that describes the observation of the soil moisture property, observed by enviot:GroveSoilMoistureSensor sensor, as shown in Figure 4.10.

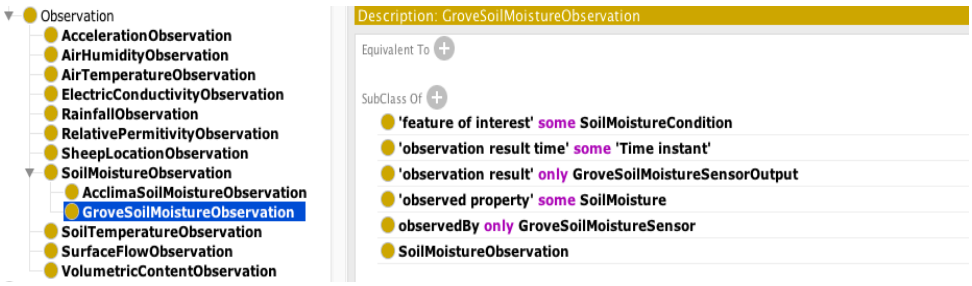


Figure 4.10: Description of enviot:GroveSoilMoistureObservation Class

### 4.5.3 The Data Module

In order to manage the data, two classes of the SSN ontology including ssn:SensorOutput and ssn:ObservationValue are extended. The output of a sensor, which is actually the result of an observation, is represented by an instance of the class ssn:SensorOutput, as shown in Figure 4.11.

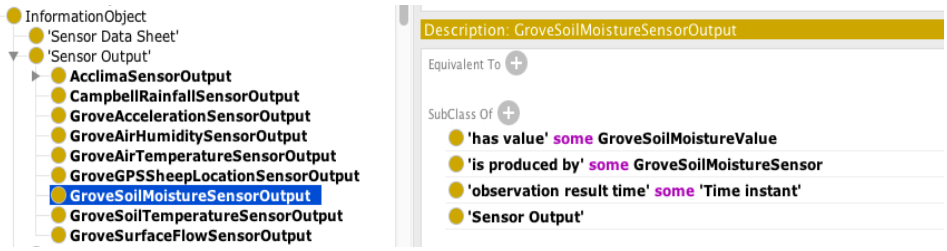


Figure 4.11: Description of the GroveSoilMoistureSensorOutput Class

The actual data value is the result of an observation, which is represented by an instance of the class ssn:ObservationValue, as shown in Figure 4.12.

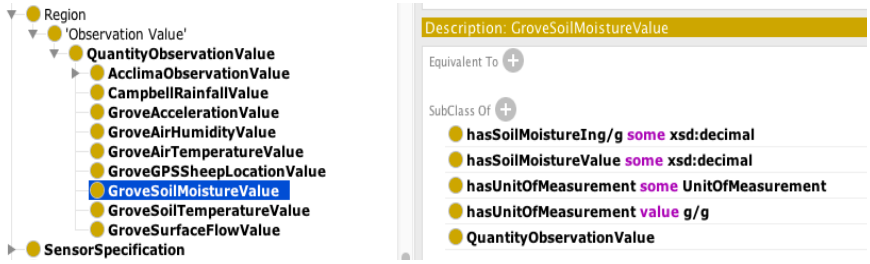


Figure 4.12: Description of the GroveSoilMoistureValue Class

### 4.5.4 The Device Module

To represent a sensor network comprising different sensors, the ssn:Device class is reused and extended, which is the sub-class of the ssn:System class. three sub-classes of the ssn:Device class are defined, i.e. enviot:SoilSensingNode, enviot:SheepTrackingNode, and enviot:WeatherMonitoring Device to describe its corresponding constituent sensors. Hence, one instance of the class enviot:SoilSensingNode would include six different sensors, i.e. GroveSoilMoistrueSensor, GroveSoilTemperatureSensor, GroveAirHumiditySensor, GroveandAirTemperatureSensor, GroveSurfaceFlowSensor, and AcclimaSensor. In order to connect a particular instance of the enviot:SoilSensingNode device with its constituent six sensors, the DUL object property , i.e. DUL:isDesbribedBy is used to point to all the constituent sensors of the device, as shown in Figure 4.13.

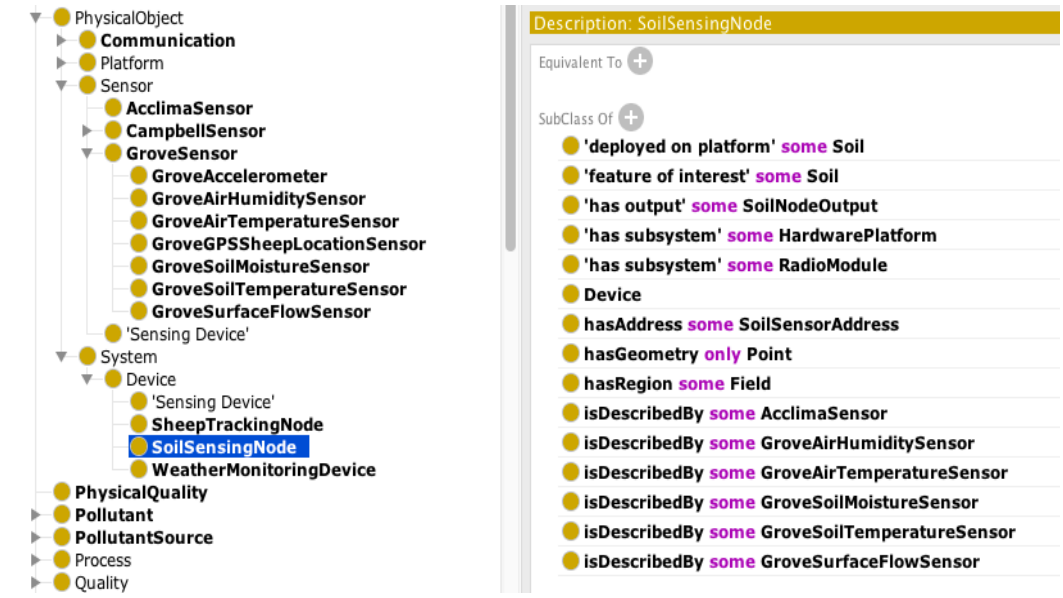


Figure 4.13: Description of the SoilSensingNode Node and its Constituent Sensors

To show the output of the devices/nodes defined in the previous step, a subclass named enviot:SoilNodeOutput of the class DUL:InformationObject is defined to represent the output of the einviot:SoilSensingNode node. To classify the output of a node/device that comprises different several sensors, a new object property named enviot:isClassifiedBy is defined to point to a particular sensor classifying the output, as shown in Figure 4.14.

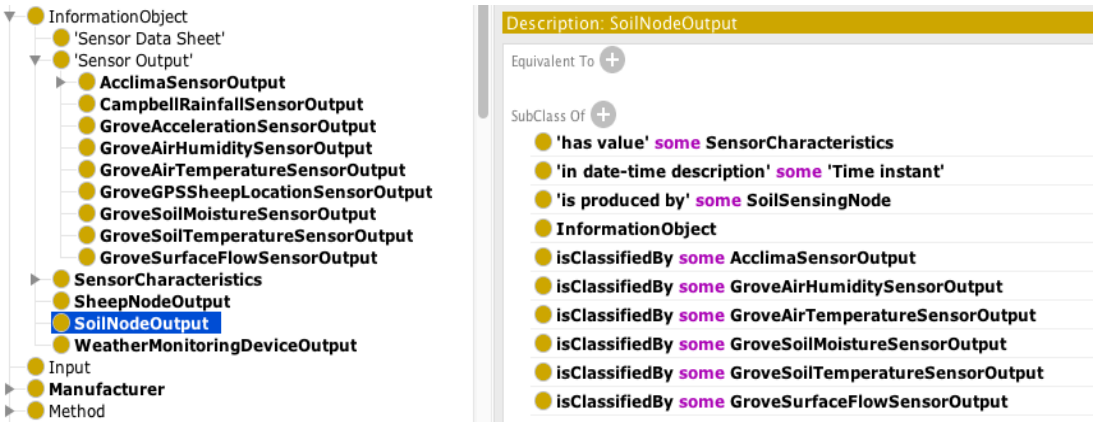


Figure 4.14: Description of the SoilNodeOutput Class

### 4.5.5 The Feature of Interest and the Property Module

Features of interest (ssn:FeatureOfInterest) are real-world entities that are defined as either events or objects and hence the target of sensing. Features of interests are

described in the SSN ontology by the ssn:FeatureOfInterest class that is defined as either DUL:Event or DUL:Object. Properties (ssn:Property) are qualities (DUL:Quality) or observable characteristics of the real-world entities (ssn:FeatureOfInterest). They do not exist independently and are the natural part of the feature of interest, for instance, in th eontology, enviot:Soil is the feature of interest and enviot:SoilMoisture is its property. The relationship between these two classes, i.e. ssn:FeatureOfInterest and ssn:Property is shown in Figure 4.15.

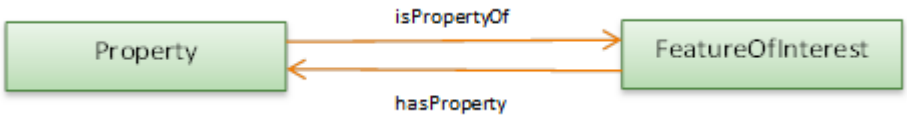


Figure 4.15: The Relationship between ssn:Propterty and ssn:FeatureOfInterest

In the ontology, several features of interest and their related properties are defined. Figure 4.16 shows the enviot:Soil as one of the features of interests and its properties pointed to by the ssn:hasProperty property.

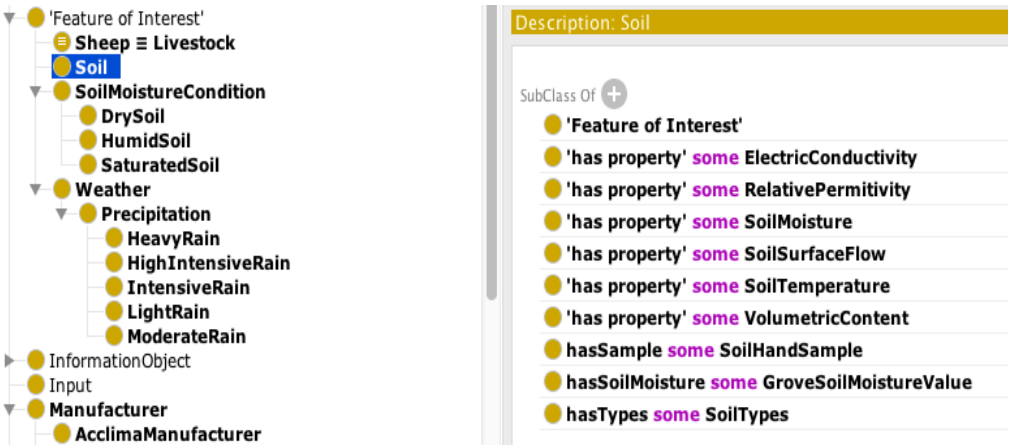


Figure 4.16: The Relationship between Feature of Interest and Property using ssn:hasProperty

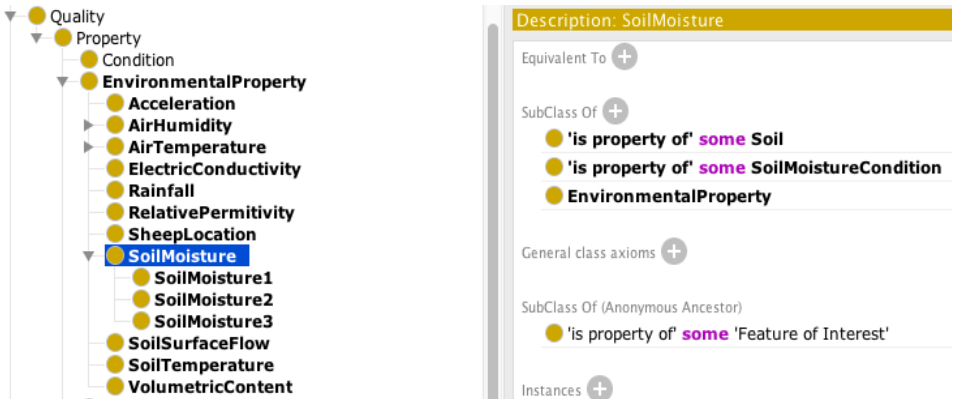


Figure 4.17: The Relationship between Feature of Interest and Property using ssn:isPropertyOf

### 4.5.6 The Geospatial Feature Module

The Basic Geo and the OGC GeoSPARQL ontologies are reused and extended in the ontology for two main purposes: to capture complex geospatial features of both the catchment area (field) and the sensor nodes deployed in it (Figure 4.6), and to track the movement of livestock in the field, for instance, whether sheep have been in the field. The geosparql:Feature class is extended and its sub-class named enviot:Field is defined to model the catchment area in the ontology. Three different types of sensor nodes in the field are deployed for sheep tracking, soil sensing and weather monitoring whose corresponsing sensor nodes are defined in the ontology as enviot:SheepTrackingNode, enviot:SoilSensingNode, and enviot:WeatherMonitoringDevice respectively. To point to the said three sensor nodes, the ssn:hasDeployment object property is used. As explained in 4.5.2 (b), the field is modelled in terms of nodes deployment into three main zones, i.e. Hilltop, Swale and Riverbank. These zones are defined as sub-classes of the enviot:Field class as enviot:Hilltop, enviot:Swale, and enviot:Riverbank, as shown in Figure 4.18.

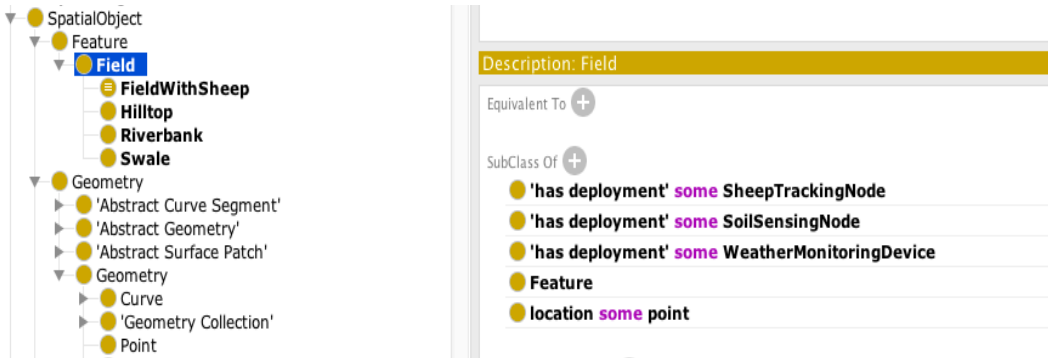


Figure 4.18: Description of the Field (Catchment area) and Its Three Zones

The instance of the class enviot:Hilltop is assigned the geometry as 'Line String' by using the geosparql:hasGeometry object property that points to the class geosparql:LineString. Similarly, the instance of the class enviot:Swale is assigned the 'Polygon' geometry that points to the class geosparql:Polygon. Finally, the instance of the class enviot:Riverbank is also assigned the 'Line String' geometry.

### 4.5.7 The Phenomenon Module

One of the main objectives of the ontology is to conceptualise different phenomena or events including risk of pollution, storm, and soil saturation, so that the ontology reasoner can infer or classify if such an event occurs. One of the major advantages of building ontologies using the OWL-DL sub-language is the automatic inferencing of class hierarchies using a reasoner. Without a reasoner, it becomes really hard to keep ontologies in a consistent and correct state. In order to classify the above said phenomena, two types of classes are defined in the ontology, i.e. Primitive Classes and Defined Classes. A primitive class, defined as a super class, is one that has only necessary conditions. Necessary conditioned are described as: if A is a member/instance of class B, then it is necessary for A to fulfil the conditions of B. Fulfilling necessary conditions alone by any instance, say C, would not make C necessarily a member of class B. In this chapter, the classes discussed so far are all primitive classes. On the other hand, a defined class is one having at least one set of both necessary and sufficient conditions. Any instance of the primitive class that also satisfies the definition of the defined class will be classified/inferred by the reasoner as an instance of the defined class. Defined classes in Protégé are distinguished from the primitive classes by having three white horizontal lines in it.

In the ontology, three phenomena are modelled and are defined as enviot:RiskOfPollution, enviot:SoilSaturation, and enviot:StormOrFloodingEvent as shown in Figure 4.19.

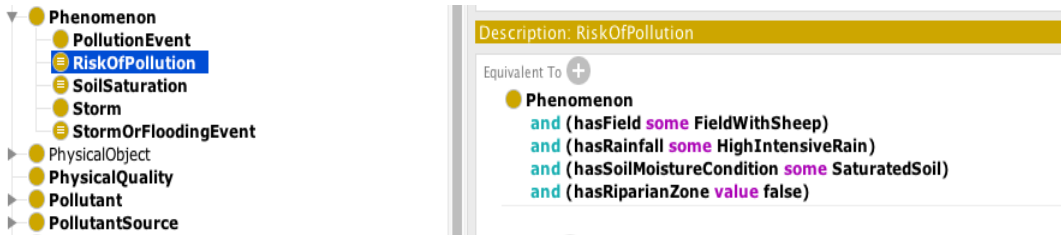


Figure 4.19: Description of the enviot:RiskOfPollution Defined Class

To check the consistency (the ontology does not include or allow for any contradictions) and functionality of these primitive and defined classes and whether the reasoner successfully infers any of the above phenomena, an instance of the class enviot:Phenomenon is created, which fulfils the definition of the defined class enviot:RiskOfPollution. After running the reasoner over the ontology, the reasoner has successfully classified (inferred) the instance of the class enviot:Phenomena under the class enviot:RiskOfPollution. Hence, this confirms the consistency and correct functionality of these classes.

### 4.5.8 The Metric Units Module

To represent quantitative measurements of the physical qualities, initially the MUO/UCUM ontology was reused and extended in the ontology. However, later this ontology was dropped for two main reasons. First, there were a lot of malfunctional xml literals leading to failure in the reasoner. The malfunctional literals were corrected, however the reasoning performance was very low. Second, the MUO/UCUM ontology populated the ontology with a large number of instances and this led to low reasoning performance. Hence, a minimal ontology is developed to overcome the above problems while describing the metric units. Two main classes are defined, i.e. the enviot:PhysicalQuality class to describe all physical qualities used in the ontology, and the enviot:UnitOfMeasurement class for representing the metric units of those physical qualities. In order to connect these two classes, two object properties are defined, i.e. enviot:MeasuresQuality and

enviot:hasUnitOfMeasurement. The domain of the enviot:MeasuresQuality is enviot:UnitOfMeasurement class and its range is enviot:PhysicalQuality class. The physical qualities are defined as the instances of the enviot:PhysicalQuality class (Figure 4.20). Then the measurement units for all physical qualities, used in the ontology framework, are defined as instances of the enviot:UnitOfMeasurement class, as shown in Figure 4.21.

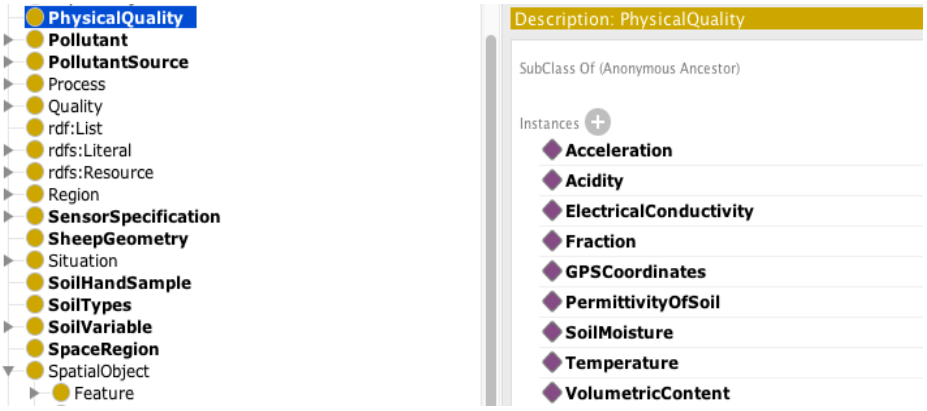


Figure 4.20: Instances of the Class enviot:PhysicalQuality

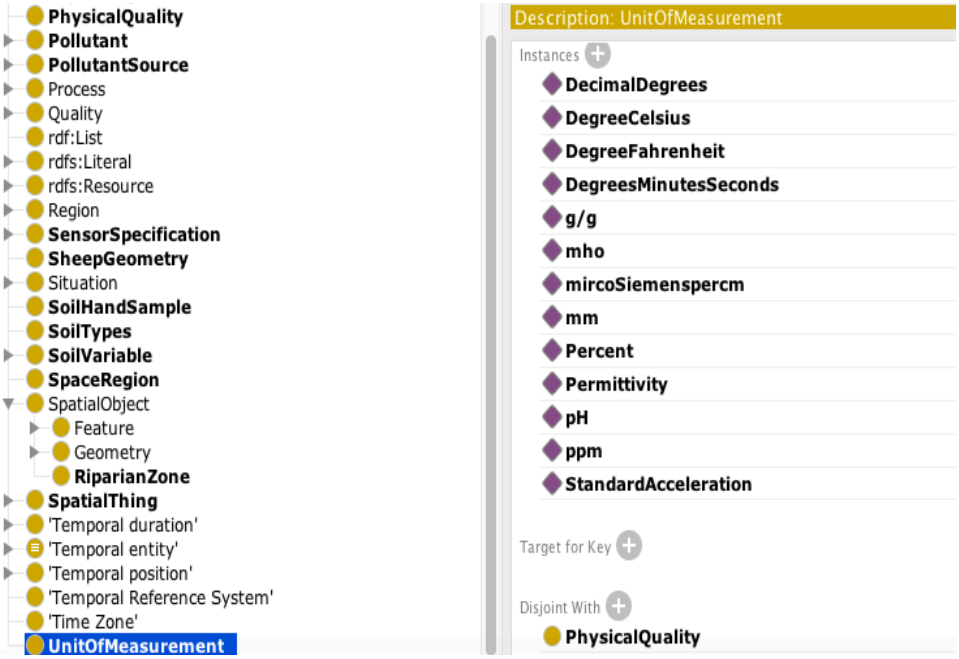


Figure 4.21: Metric Units Defined by the enviot:UnitOfMeasurement Class

## 4.6 Summary

The final generic block diagram and a sample of the ontology are shown in Figure 4.22 and **Figure 4.23** respectively. The ontology has integrated and extended existing standard minimal domain and upper ontologies. The core of the ontology is the extended versions of these standard ontologies including SSN, GeoSPARQL, Time, and Geo, coupled with the applicaton ontology for describing streaming data derived from the Environmental IoT infrastructure. However, the ontology has not used any task/method ontology that usually focuses on the problem-solving domains to accomplish a particular goal, for instance, expert systems, which is not the purpose of this work. Furthermore, the MUO/UCUM has been removed owing to the issues of having malfunctional xml literals and a large number of individuals/instances, leading to both reasoning failure and low performance. Hence, a minimal lightweight unit ontoloy is devloped. The consistency of the classes in the ontology has been checked successfully using the Pellet reasoner. The ontology owl file can be downloaded from the link given below.

https://lancaster.box.com/s/zhbwidfd03gulhqtkojgebsfrdyza8i8

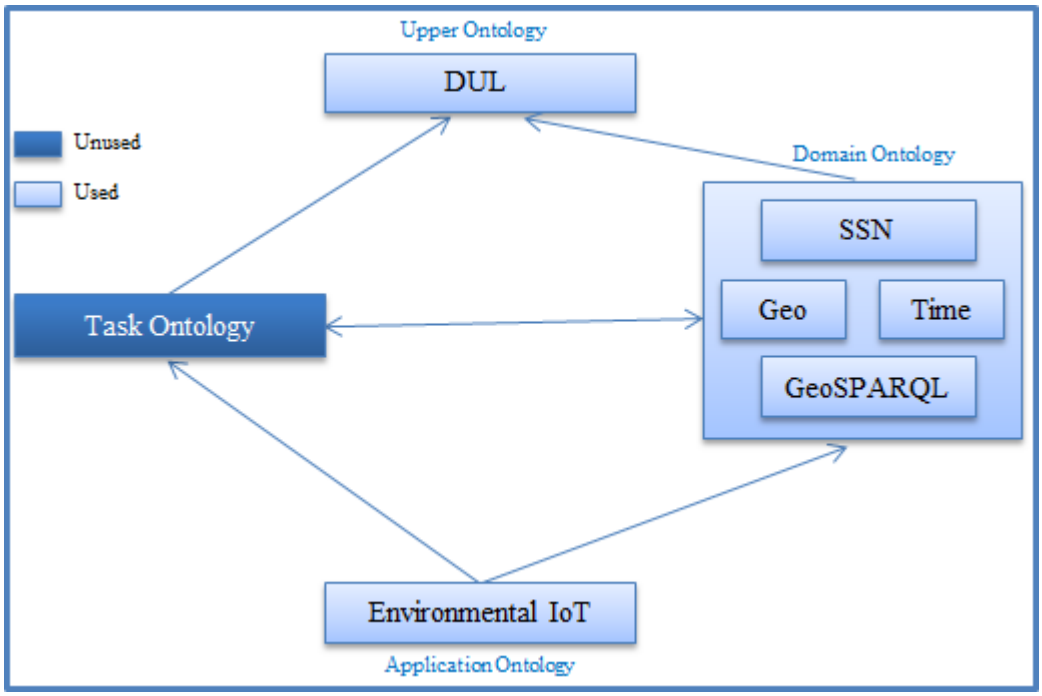


Figure 4.22: Environmental IoT Ontology

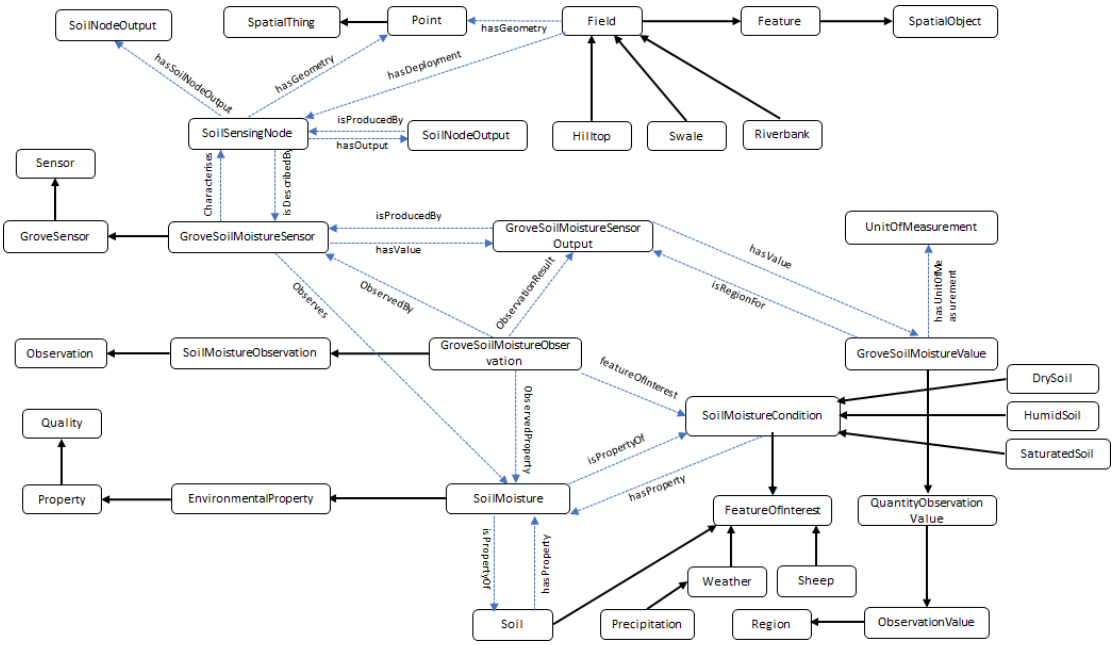


Figure 4.23 A sample diagram of the ontology. The rectangular boxes represent classes/concepts, the solid lines (linking a class to another class) represent rdfs:subClassOf relations and the dashed labelled lines represent the object properties.

## 4.7 Conclusion

This chapter has discussed in depth the design and development of an integrated ontology for the semantic enrichment of environmental data stemmed from the Environmental IoT infrastructure deployed in the natural environment. The ontology has taken into account various key design criteria including reusing existing standards, modularity and extensibility, expressiveness and reasoning support and aiming for a lightweight design. The chapter has discussed the core modules of the ontology focussing on three main themes of sensor metadata representation, i.e. thematic, spatial and temporal. Though the ontology has not demonstrated yet the underlying functionality, strengths and limitations in the target domain, it is checked against anomalies and inconsistencies using the Pellet reasoner and has found consistent. No inconsistencies have been detected in the ontology itself (super-class/sub-class relationships), or in the set of individuals (instances) of the classes that have been defined to test the working of the ontology. The domain and range definitions have found compatible, cardinality properties are consistent and the requirements on properties' values do not conflict with domain and range restrictions. The ontology does not conflate observational data with the properties of sensors

which was one of the limitations of CESN [160] ontology potentially leading to data integration issues. The ontology has described the knowledge in the target domain along with space and time concepts, covering all thematic, spatial and temporal dimensions which have not found in the work of [163]. The ontology has extended existing standard ontologies leading to better semantic interoperability support contrary to the approach in [129] using O & M and SensorML specification which lacks explicit semantic interoperability. Moreover, the ontology provides strong querying support that was lacked in the approach used in [168].

The next chapter provides an evaluation of the ontology through three different real-world use-cases, derived from the analysis of the semi-structured interviews and IoT project meetings with environmental scientists.

# 5 Evaluation

This chapter provides an evaluation of the ontology through three different real-world use-cases, derived from the analysis of the semi-structured interviews, and the Environmental IoT project meetings with environmental scientists. These use-cases are based on near real-time data stemming from the Environmental IoT Infrastructure [20]. The ontology design and evaluation are intrinsically linked through the iterative approach as introduced in Chapter 4.

The rest of the chapter is structured as follows. Section 5.1 describes three real-world use-cases, which are the risk of a pollution event, geospatial data integration and reasoning, and interoperable metric units. Section 5.2 looks more closely at the framework that has been set up to carry out the evaluation of the use-cases. In section 5.3, the evaluation criterion is briefly described. Section 5.4 and 5.5 then provide a more in-depth assessment of the work through evaluation of the use-cases and addresses the research questions associated with the aims of this thesis. Section 5.6 discusses the analysis and the lessons learned from this work. Finally, section 5.7 presents concluding remarks.

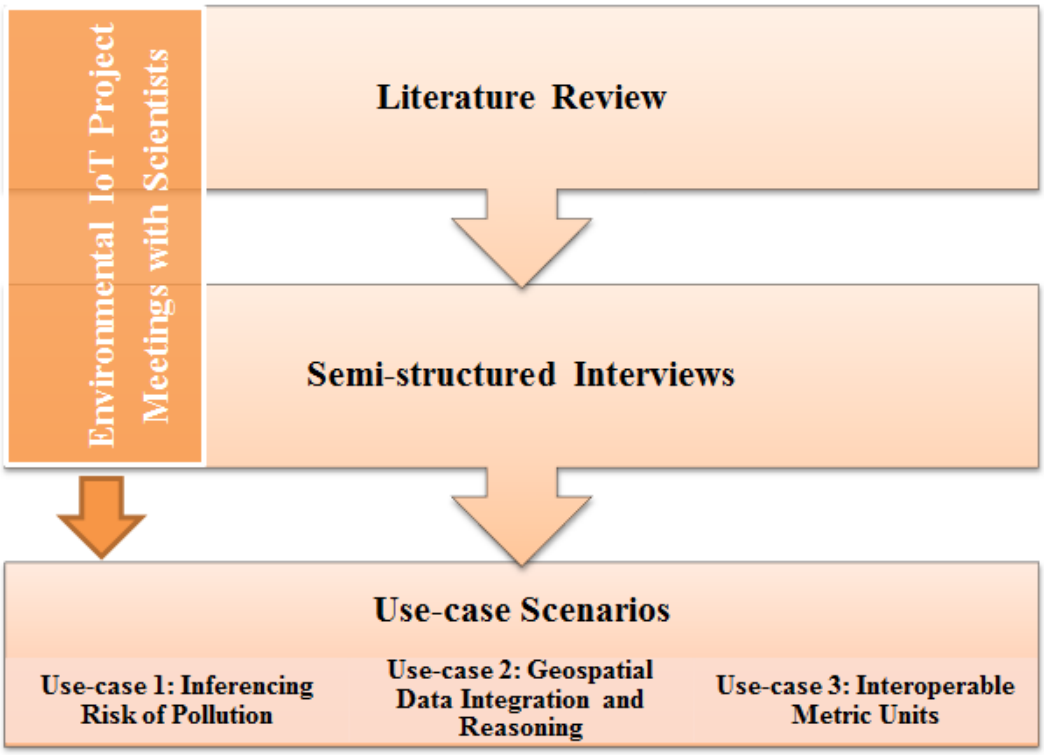


Figure 5.1: Use-cases Derivation

## 5.1 Real-world Use-cases

This section describes the three use-case scenarios derived from the semi-structured interviews and drawing upon the main key findings, and the Environmental IoT project meetings with environmental scientists, as shown in Figure 5.1. These use-cases are described below.

### 5.1.1 Use-case 1: Risk of Pollution Event

Sara is a senior soil scientist at the Centre of Ecology and Hydrology (CEH) whose areas of research are soil, biogeochemical and ecosystem science. She is also very interested in knowledge systems exploiting advances in computer science. She investigates the impact of land management on ecosystem services, change in soil structure, and impacts of nitrogen pollution on soils. Currently, she has been working on a research project with her colleague George, a hydrologist, investigating the interdependencies among different environmental facets such as soil, livestock movement, weather, chemical fertilisers and water quality and their impact on each other. Their focus is on one specific geographic region around the Conwy in North

Wales, typical of many rural areas supporting important industries including agriculture, forestry, tourism and fishing. They want to identify the potential anomalous events regarding pollution which may occur, for instance, the movement of livestock into lowland areas, combined with intensive rainfall, can cause a significant transfer of nutrients and faecal bacteria into coastal waters. Their research questions are: Is there a risk of occurrence of pollutants in water? If yes, what could be the cause of the pollution that occurred in water? They believe that the usage of Internet of Things technology along with techniques based on richer knowledge-driven use of data would possibly help in predicting the occurrence of these events (with a sample reasoning framework shown below in Figure 5.2).

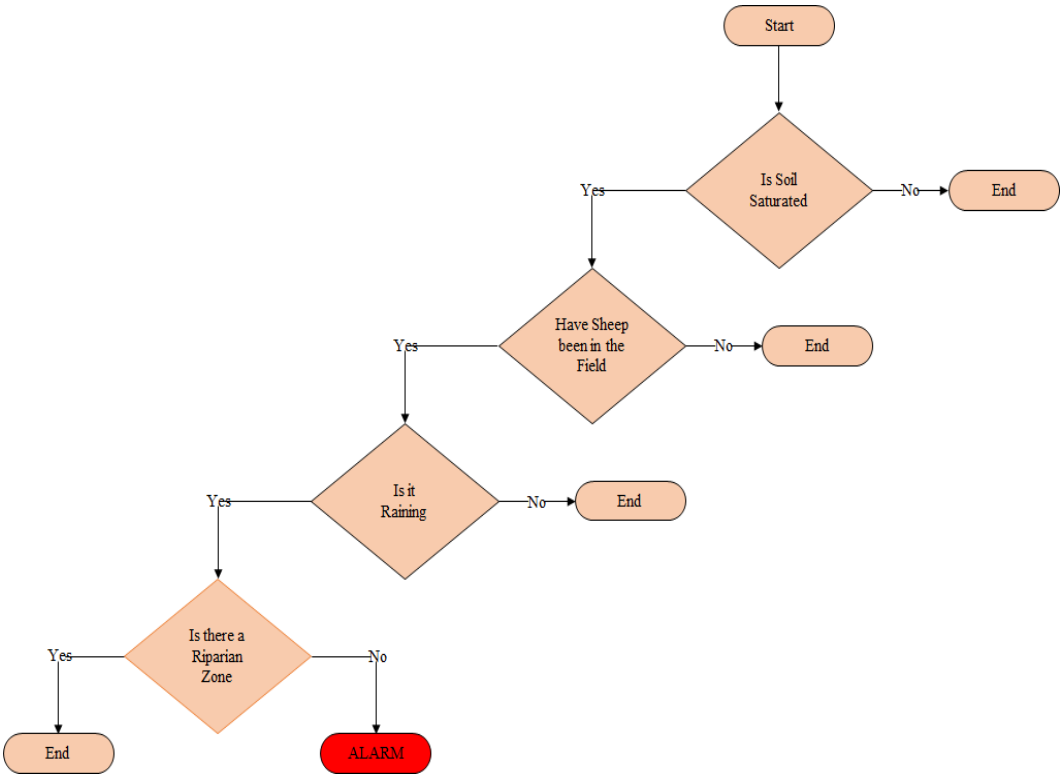


Figure 5.2: The Use-case of a Potential Risk of Pollution Event in the Catchment

### 5.1.2 Use-case 2: Geospatial Data Integration and Reasoning

Sara and George want to discover and understand the spatio-temporal trends in their catchment area in order to be able to respond to such emerging trends or geographic events in a timely manner. To do so, they are trying to understand the geospatial and temporal dimensions of several environmental variables and need to integrate these observations collected from the sensors. They want to merge these measurements

captured at different locations and times to get a unified view of the data and an understanding of how different environmental variables are related to one another. They require a richer knowledge-driven database and smart data retrieval techniques to manage their data more effectively and to add meaning to their metadata that the traditional database management systems they previously used cannot do. They also need to find the answers of various kinds of complex queries anytime they want to retrieve from the smart knowledge base, for instance, what are the features and geometries of the catchment where sensors are deployed, where exactly the sheep have been found in the field (e.g. hilltop, swale, and riverbank), what was the soil moisture value when the storm, say, Storm Desmond, occurred and how long the flood did last for, what is the location of soil sensing node where the soil has been saturated, what types of and how many sensors are deployed on the river bank measuring soil moisture, soil temperature, and sheep movement and where are the high risk pollutants' zones etc.

### 5.1.3 Use-case 3: Interoperable Metric Units

Sara and George have been collecting different measurements from a sensor network deployed in the catchment such as soil moisture and temperature, electric conductivity and permittivity of soil, air humidity and temperature, cattle movement in the field, rainfall measurements, flow detection of water etc. They have also got some data regarding soil nutrients and pH from hand sampling method and analysing it in the lab. In order to provide meaning to these quantitative measurements they are using several metric units. The collected data is sent to the cloud-based server in a remote site via GSM for storage and further processing where it has also been accessed and used by their collaborators working in the same or connected area of environmental science. Sara, being a soil scientist, is interested in one aspect of the data and is using her own chosen metric units, George, being a hydrologist, is interested in another aspect of the data and is representing it with different units while their collaborators, working in soil science, hydrology and biogeochemistry, are looking for different aspects of the same data with different metric units. This situation is exacerbated when all these scientists and modellers do not follow uniform metric units. Using different metric units for the same physical qualities has arisen the issue of heterogeneity. To cope with this metric unit conflict, they have decided to use

common measurement units e.g. SI units but this approach is not viable in situation when they need an automated integration of different measurement datasets from these various connected subdomains so there is a need to recognise such heterogeneity in metric units and manage the consequent need for interoperability in a more automated manner.

## 5.2 Evaluation Framework

A framework has been set up to evaluate the real-world use-cases. Figure 5.3 describes the steps involved in the iterative approach. The data is stemming from the Environmental IoT Infrastructure, deployed in the natural environment to monitor different environmental facets. Data is collected in JSON format from the IoT devices and is stored in a MongoDB NoSQL database installed on a cloud server. The JSON format is adopted because of its advantages including ease of use, compatibility and lightweight syntax. The data is then semantically enriched with the vocabularies of the ontology designed and developed in the previous chapter. The Python scripts are written by one of the project collaborators to perform data transformation (RDFization). The result of the semantic enrichment is the JSON-LD data that is one of the serialisations of the RDF. These JSON-LD triples are loaded in the GraphDB triplestore for processing. The triples along with the ontology are fed into a Jena application framework to perform the intended tasks including deducing new knowledge by using its inference engine (e.g. deriving the pollution event), geospatial data integration and interoperability. If the target tasks are not accomplished, the process goes back to the ontology development phase where the ontology is modified (according to the iterative approach described in Chapter 4). The data is then semantically enriched with the modified ontology to reflect the changes. The newly transformed data along with the modified ontology are fed back into the Jena application. Hence, both the ontology and the semantically enriched sensor measurements are refined in every iteration until the intended tasks are done.

As can be seen in Figure 5.3, there are four main tasks: i) ontology refinement that is done after every iteration; ii) semantic enrichment of data and then converting data to JSON-LD serialisation (an example of JSON-LD data after semantic enrichment of one particular sample of soil sensing node can be seen in Figure 5.4); iii) storing and retrieving RDF triples in the GraphDB [197] triplestore (GraphDB is a highly

efficient and robust graph database with RDF, SPARQL and GeoSPARQL ontology support and the advantages of GraphDB over other triplestores are its compliance to W3C standards and support for highly efficient reasoning); iv) access to GraphDB through the Jena application framework using an API (Jena is an open source java-based application framework for building Semantic Web applications, which provides a programming interface for RDF, OWL and SPARQL). Jena also includes a rule-based inference engine. In Semantic Web approaches, inference is used to deduce further knowledge based on existing RDF triples and a set of inference rules using an inference engine (reasoner). In this work, a Jena application has been developed, which takes RDF triples as an input along with a set of rules and the ontology to infer new knowledge.

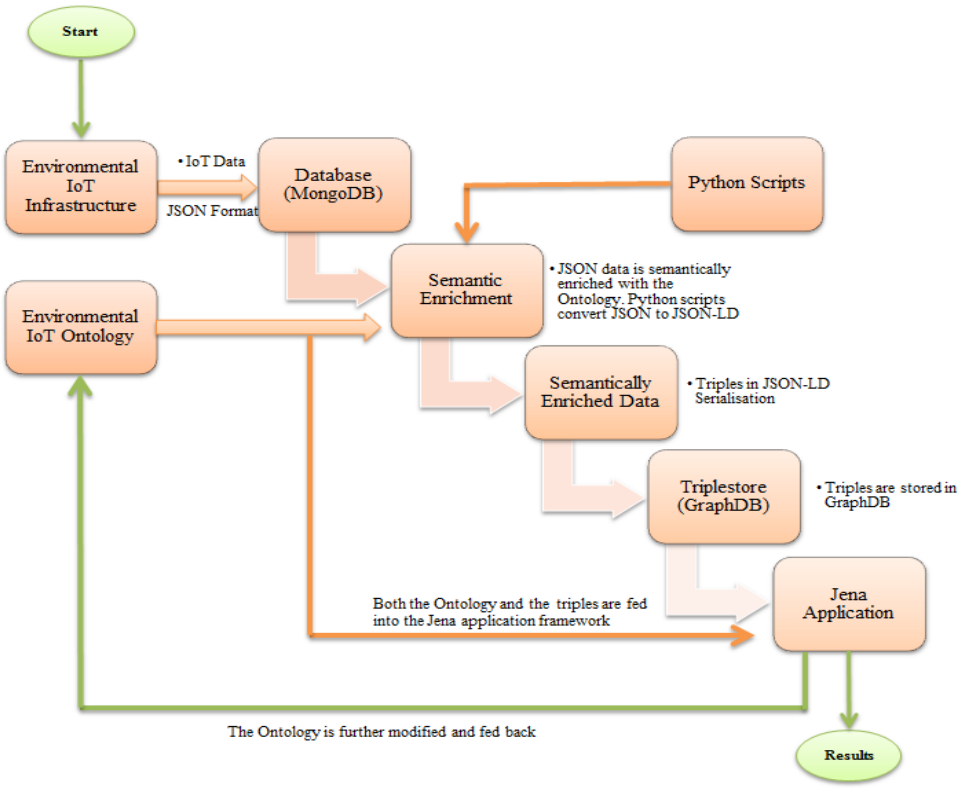


Figure 5.3: Evaluation Framework of the Overall Approach

{"_id":"55fbf0a270b2d7302700002a","rdf:type":"enviot:SoilSensingNode","geosparql:hasGeometry":{"sf:asWKT":"POINT(-3.133015 53.225369)","@id":"sf:Point","xsd:type":"geosparql:wktLiteral"},
"ssn:hasSubSystem":{"rdfs:subClass":"enviot:HardwarePlatform","@id":"enviot:ArduinoMega2560"},"ssn:deployedOnPlatform":{"rdf:type":"ssn:Platform","@id":"enviot:Soil"},"enviot:hasAddress":"\"A1\"","DUL:isDescribedBy":[{"ssn:hasMeasurementCapability":"enviot:GroveSoilMoistuMeasurementCapability","rdfs:subClass":"ssn:Sensor","@id":"enviot:GroveSoilMoistureSensor"},{"ssn:hasMeasurementCapability":"enviot:GroveAirTemperatuMeasurementCapability","rdfs:subClass":"ssn:Sensor","@id":"enviot:GroveAirTemperatureSensor"},{"ssn:hasMeasurementCapability":"enviot:GroveSoilTemperatuMeasurementCapability","rdfs:subClass":"ssn:Sensor","@id":"enviot:GroveSoilTemperatureSensor"},{"ssn:hasMeasurementCapability":"enviot:GroveAirHumidityMeasurementCapability","rdfs:subClass":"ssn:Sensor","@id":"enviot:GroveAirHumiditySensor"}],

"@context":{"owl":"http://www.w3.org/2002/07/owl#","wgs84_pos":"http://www.w3.org/2003/01/geo/wgs84_pos#","DUL":"http://www.loa-cnr.it/ontologies/DUL.owl#","xsd":http://www.w3.org/2001/XMLSchema#
,"geosparql":"http://www.opengis.net/ont/geosparql#","rdf":"http://www.w3.org/1999/02/22-rdf-syntax-ns#","rdfs":"http://www.w3.org/2000/01/rdf-schema#","sf":"http://www.opengis.net/ont/sf#",
"ssn":"http://purl.oclc.org/NET/ssnx/ssn#","enviot":"http://www.environmental-iot.com/enviot_ontology/IotSemanticModel#","time":"http://www.w3.org/2006/time#","enviot:TimeStamp":{"enviot:hasQuantityValue":{"@type":"xsd:dateTime"},"time:inDateTime":{"time:inDateTime":"time:DateTimeDescription","@id":"time:Instant"}}},"DUL:hasLocation":{"rdf:type":"DUL:SpaceRegion","@id":"enviot:Hiraethlyn"},"@id":"55fbf0a270b2d7302700002a","@type":"enviot:SoilSensingNode"}

Figure 5.4: JSON-LD Representation of One Particular Soil Sensing Node

## 5.3 Evaluation Criteria

From Figure 5.3, it can be seen that the evaluation framework revolves around the ontology that is designed and developed on top of the Environmental IoT Infrastructure (see Chapter 4). To remind the reader, the main goal of the ontology is to accomplish three main tasks, i.e. discovering the interdependencies across disparate datasets, spatio-temporal data integration and reasoning and metric units interoperability. To achieve these objectives, the ontology is plugged into an application built in the Jena programming framework. As the ontology is a major component of this application framework, the evaluation of the results of the application is mainly dependent on the ontology but also partly on the application framework.

The evaluation criterion is based on the following key reflective qualitative aspects.

- Functional – accomplishing the above said three key tasks that are important to the data needs of the Environmental IoT Infrastructure.
- Expressive – how well does the approach do the job, e.g. is a query natural for an environmental scientist to write?
- The overall strengths and limitations of the approach.

In summary, the evaluation is carried out to evaluate the applicability, strengths and limitations of the adopted approach in the target domain.

## 5.4 Use-cases Evaluation

This section describes the in-depth evaluation of the overall approach through the above real-world use-cases. Use-cases are based on sensor measurements that are semantically enriched with the vocabulary of the ontology. This semantically enriched data is intended to provide support to achieve the objectives of environmental scientists (e.g. Sara and George, mentioned in the use-cases) regarding environmental data. More specifically, the approach is designed to enable Sara and George: to discover the interdependencies between disparate datasets representing different environmental facets, to answer their complex geospatial queries by integrating their datasets in a unified way, and to provide unambiguous automated interoperability between different metric units. All these use-cases are briefly evaluated one by one, with an overall evaluation then carried out across all the use-cases and against the above criteria.

### 5.4.1 Evaluating Use-case 1: Risk of Pollution Event

To infer the risk of pollution event in the catchment, first it is important to identify a pollution event and the scenario through which it occurs. In the proposed approach, the pollution event is conceptualised in the ontology (see section 4.5.7). The description of the scenario for the pollution event is derived from Figure 5.2, which is defined by environmental scientists who collaborated in the project. From the figure, it can be seen that deducing the risk of a pollution event is a complex event which further depends on the knowledge about soil saturation, high intensive rainfall, existence of sheep in the field and existence of any riparian zones. Deriving such a complex event would require further support from the OWL language. Though OWL-DL provides considerable expressive power, it has some limitations, particularly in terms of describing properties and individuals [198]. In order to overcome this limitation, SWRL (Semantic Web Rule Language) [199] rules are defined to provide additional expressive power. SWRL is an expressive OWL-based rule language, which allows users to define rules in the ontology providing powerful deductive reasoning capabilities [200]. Hence, to deduce the potential pollution event in the

ontology, initially SWRL rules are defined which are based on the conditions described in Figure 5.2.

**SWRL Rules**

A SWRL rule contains an antecedent (body) and a consequent (head), each of which is formed from a set of atoms, and has the form: antecedent => consequent

A SWRL rule can be read as if the antecedent is true, then the consequent must also be true. Using atoms in a SWRL rule would become:

atom ^ atom => atom ^ atom

The above rule can be read as if all the atoms in the antecedent are true, then the atoms in the consequent must also be true. An atom is an expression of the form:

D (x), P (x, y), or built-in (r, x,…)

In the above expression, D is an OWL description or data range, P is an OWL property, r is a built-in relation, x and y are either variables, OWL individuals, or OWL data values [201]. Atoms can represent classes, instances, data literals, individual variables or data variables. All variables are preceded by a question mark (?) in the rule.

The following four SWRL rules are defined in the ontology to deduce the risk of a pollution event. Rule 1 is defined to infer the high intensive rainfall that is modelled in the ontology. An instance of the class enviot:Weather is created with the name 'WeatherRainfall' and is assigned a rainfall value '100'. In the ontology, a sub-class, called enviot:HighIntensiveRain, of the class enviot:Weather is defined. An assertion is put on the enviot:Weather class, which defines the range of the rainfall values, as shown in Figure 5.9. After modelling the required information, the rule is defined as:

**Rule 1: Soil Saturation**

Rule 1 is defined to infer a soil saturation event that is modelled in the ontology as a sub-class of the enviot:Phenomenon class. In the ontology, three types of soil moisture conditions are described which characterise the current status of soil: i) dry soil is the one whose soil moisture value lies in the range 0-299; ii) humid soil's

moisture value falls in the range 300-599; iii) saturated soil is one whose soil moisture value is measured in the range 600-900. If any sensor measurement of soil moisture falls in this third category, the reasoner would classify the status of soil as saturated. To derive a soil saturation event, rule 1 is defined as:

enviot:Soil(?s) ^ enviot:hasSoil(enviot:Hiraethlyn, ?s) ^ enviot:hasSoilMoisture(?s, ?m) ^ enviot:hasSoilMoistureValue(?m, 900) -> enviot:SoilSaturation(?s)

After adding the above rule to the ontology, the Pellet reasoner is selected and started to reason over the ontology. The output of the reasoner is shown in Figure 5.5. From the description view in the figure, it can be seen that an individual named 'Soil1' (shaded in yellow), which is an individual of the class enviot:Soil, is correctly classified as an individual of the class enviot:SoilSaturation.

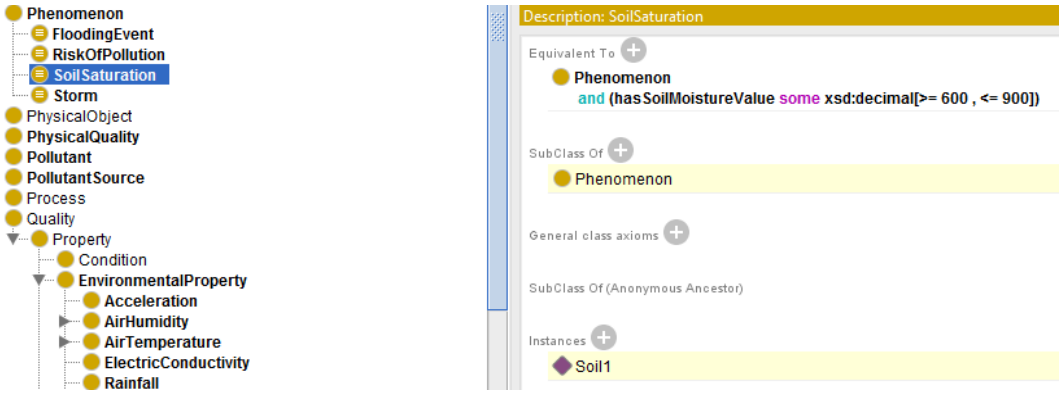


Figure 5.5: Individual (Soil1) Classified as being an Individual of enviot:SoilSaturation

**Rule 2: Sheep in the Field**

In order to derive whether sheep are found in the field, a sub-class named enviot:FieldWithSheep of the class enviot:Field is defined. Note that field is situated in the region named Hiraethlyn. An individual of the class enviot:Field is defined having an object property enviot:hasSheep, that would point to sheep in that field. If that individual has got sheep located with it, it would be classified as being an individual of the class enviot:FieldWithSheep. The rule is defined as:

enviot:Field(?f)^enviot:Sheep(?s)^enviot:hasField(enviot:Hiraethlyn,?f)^enviot:hasSheep(?f, ?s) -> enviot:FieldWithSheep(?f)

The result of the reasoning is shown in Figure 5.6. From the figure, it can be seen that the individual named SheepyField1 is correctly classified as the individual of the class enviot:FieldWithSheep.

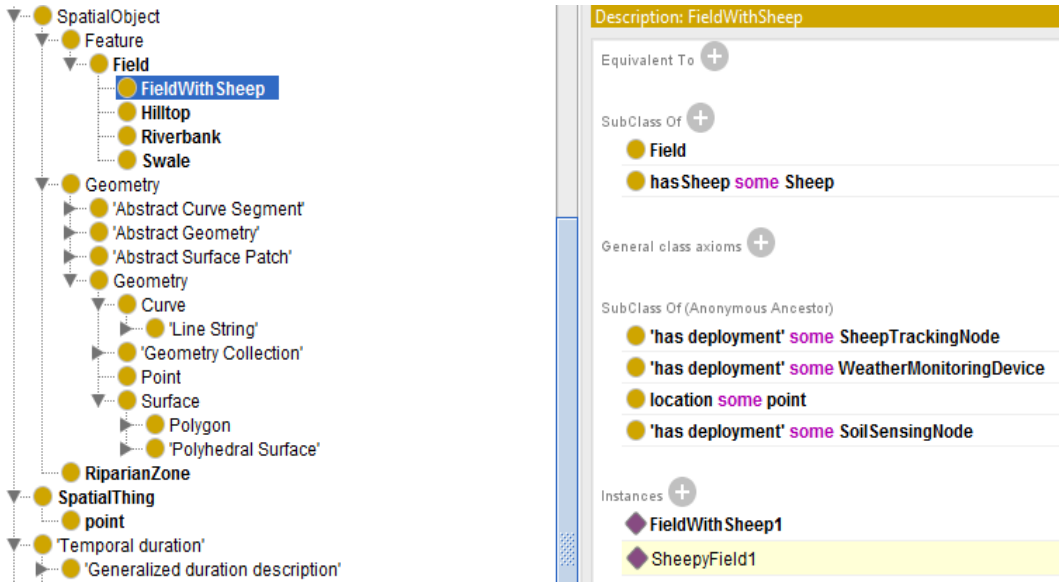


Figure 5.6: Individual (SheepyField1) Classified as being an Individual of the class enviot:FieldWithSheep

**Rule 3: High Intensive Rainfall**

To infer whether there is high intensive rainfall, different types of precipitation are modelled in the ontology with particular rainfall measurement ranges. These types include light rain, moderate rain, heavy rain, intensive rain and high intensive rain. All these types of rain are assigned value ranges. To define rainfall ranges for these different types of rain, data type restrictions are used. The rainfall range for high intensive rain is defined between 50 and 400 inclusive. An individual of the class enviot:Weather is created in the ontology and is assigned an arbitrary rainfall value 100. This value is compared against the rainfall value and falls in the category of enviot:HighIntensiveRain. The rule is defined as:

enviot:Weather (?w) ^enviot:hasWeather (enviot:Hiraethlyn, ?w) ^ enviot:hasRainfall (?w, ?r) ^ enviot:hasRainfallValue (?r, 100.0) -> enviot:HighIntensiveRain (?w)

After adding the above rule to the ontology, the Pellet reasoner is invoked to infer high intensive rain if the conditions in the antecedent are true. The result of the reasoning process is shown in Figure 5.7. From the figure, the individual (WeatherRainfall shaded in yellow) that is a member of the class enviot:Weather is inferred as an individual of the class enviot:HighIntensiveRain, which confirms that the rainfall event is modelled in the ontology accurately.

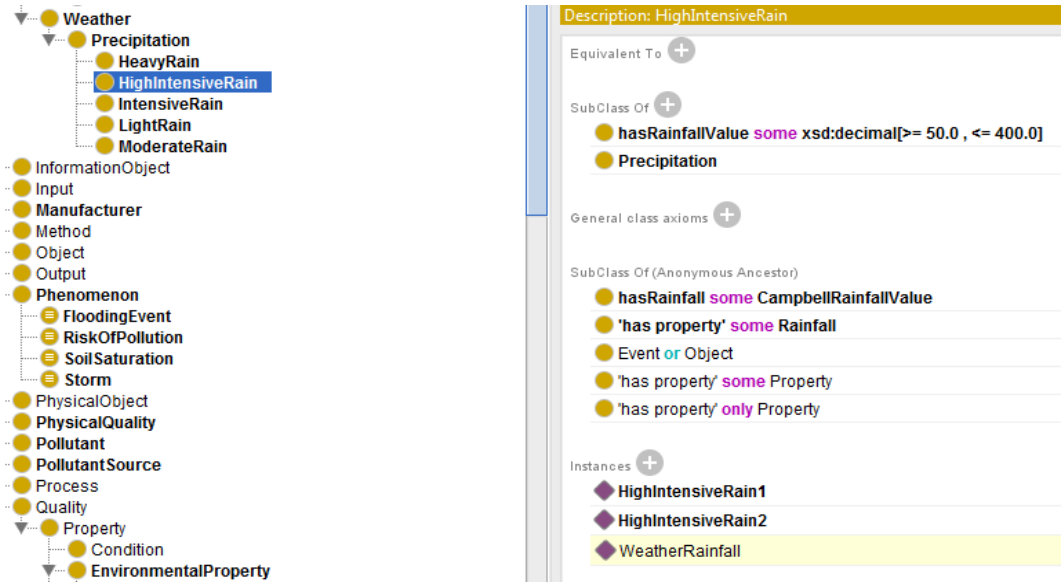


Figure 5.7: Classification of WeatherRainfall as a High Intensive Rain

**Rule 4: Risk of Pollution**

Finally, rule 4 is defined to infer the risk of a pollution event if all the above rules are met. Hence, this rule checks if the above three events are met along with a Boolean data type enviot:hasRiparianZone that must be false. An individual of the class enviot:Phenomenon is created having three property restrictions regarding the above three sub-events. If the conditions in the antecedent are met, the individual will be classified as the member of the class enviot:RiskOfPollution. To deduce this event, the SWRL rule is defined as:

enviot:Phenomenon(?p)^enviot:hasField(?p,enviot:FieldWithSheep1)^enviot:hasRainfall(?p,enviot:HighIntensiveRain1)^enviot:hasSoilMoistureCondition(?p,enviot:SaturatedSoil1) ^ enviot:hasRiparianZone(?p, false) => enviot:RiskOfPollution(?p)

After running the Pellet reasoner over the ontology, the atoms in antecedent in Rule 4 are found to be true, hence the atoms in the consequent are fired. Consequently, the reasoner classifies the individual, named PhenomenonPollution, of the class enviot:Phenomenon as the individual of the class enviot:RiskOfPollution. Figure 5.8 illustrates the result of the reasoner. The inferred individual PhenomenonPollution is shaded in yellow.

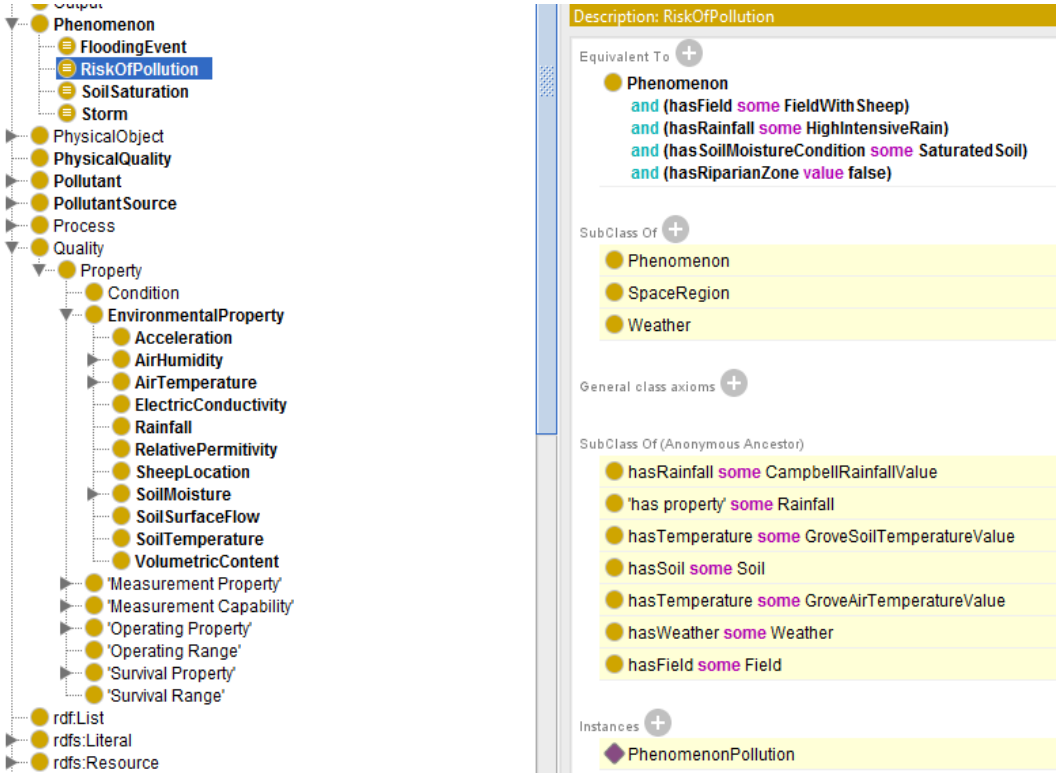


Figure 5.8: Illustration of Inference of the Risk of Pollution Event

From the evaluation above, it can be seen that using SWRL rules has two main advantages: i) SWRL extends the expressiveness power of OWL in a simple way allowing us to check the consistency of the classes and the inference of new knowledge; ii) SWRL is compatible with OWL syntax and semantics. However, this approach has two main limitations: i) the additional expressive power that comes from the SWRL rules leads to inefficiency in reasoning; ii) SWRL does not have a mechanism to access external data sources, thus requiring all the data to be brought

into the ontology. This could lead to a huge overhead. Hence, there is a need of more flexible and efficient rule-based approach that could deduce new knowledge while not affecting the reasoning efficiency.

**Jena Inference Rules**

In order to overcome the limitations of SWRL rules, a set of inference rules is defined and implemented in the Jena application on top of the ontology and RDF triples using the general-purpose rule-based reasoner [202]. The set of rules along with the ontology and RDF triples (JSON-LD files) are given as an input to the Jena application. The application uses the in-built general-purpose rule engine and infers new facts if the rules are triggered ('fired') successfully. Like SWRL, the rules in Jena follow the same antecedent -> Consequent form, however the syntax is based on SPARQL. The inference rules in Jena are defined below.

Rule 1 is about inferring the soil saturation event. Environmental scientists have calibrated soil moisture sensors for three different ranges of soil moisture values, as described above which are 'Dry Soil', 'Humid Soil', and 'Saturated Soil'. All these concepts have been specified in the ontology. The property restriction range of enviot:SaturatedSoil is shown in Figure 5.9.

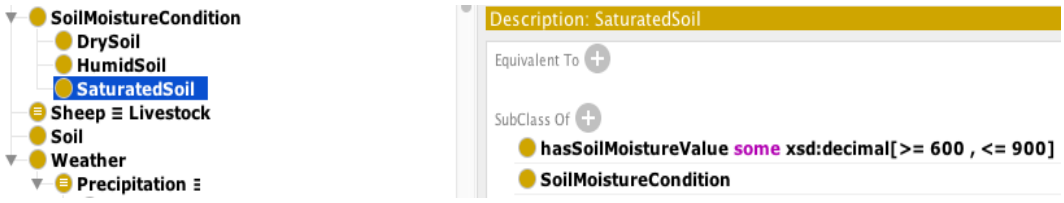


Figure 5.9: Description of the class enviot:SaturatedSoil

If the soil moisture value, collected by a particular soil moisture sen`sor, exceeds 599, the Jena reasoner should classify the soil moisture condition as 'Saturated Soil'. The inference rule to derive this event is defined below.

[**rule1SoilSaturation**:(?s rdf:type enviot:SoilNodeOutput), (?s enviot:isClassifiedBy ?senout),(?senout rdf:type enviot:GroveSoilMoistureSensorOutput), (?senout ssn:hasValue ?gmval), (?gmval rdf:type enviot:GroveSoilMoistureValue), (?gmval enviot:hasSoilMoistureValue ?val), ge(?val, 600), le(?val, 900) -> print (?val,'Soil is Saturated')]

In the above rule, ‘s’ is an instance of type output which is produced by the soil sensing node. This output is classified by a particular sensor output that has a particular soil moisture quantity value. If that value exceeds 599 then the above rule should ‘fire’ and derive that fact that soil is saturated. The result of the above rule is shown below:

```
'825'^^http://www.w3.org/2001/XMLSchema#integer <’Soil> <is> <Saturated’>
'612'^^http://www.w3.org/2001/XMLSchema#integer <’Soil> <is> <Saturated’>
'742'^^http://www.w3.org/2001/XMLSchema#integer <’Soil> <is> <Saturated’>
```

For the sake of simplicity, the URIs of the above triples are not shown in the results. It can be seen from the above results that only those instances are inferred whose soil moisture values are greater than 599 and less than 900, which indicate that soil has been saturated. The soil saturation phenomenon is important for soil scientists because it gives an indication of one of the factors of risk of pollution. When this condition is satisfied, scientists want to know more about the rainfall measurements in the catchment.

The second rule infers the high intensive rainfall event, which is defined as:

[**rule2HighIntensiveRainfall**: (?w rdf:type enviot:WeatherMonitoringDeviceOutput), (?w enviot:isClassifiedBy ?senout), (?senout rdf:type enviot:CampbellRainfallSensorOutput), (?senout ssn:hasValue ?crval), (?crval rdf:type enviot:CampbellRainfallValue), (?crval enviot:hasRainfallValue ?val), (?hir rdf:type enviot:HighIntensiveRain), (?hir enviot:hasRainfallValue ?hirval) , ge (?val, ?hirval) -> print (?val, ‘Rain is High Intensive’)]

In the above rule, ‘w’ is defined as an individual of the device output, which is generated by a weather monitoring device. This output is classified by a rainfall sensor output generated by a particular rainfall sensor. The sensor output has a particular rainfall value which is compared with the measurements of high intensive rainfall ranges described in the ontology (Figure 5.10). If the rainfall value is greater or equal than the range of the high intensive rain value, it should be classified as high intensive rain.


Figure 5.10: Description of the class enviot:HighIntensiveRain

The result of the above Jena inference rule is shown below:

```
'51.0'^^http://www.w3.org/2001/XMLSchema#decimal <‘Rain> <is> <High>
<Intensive’>
'60.0'^^http://www.w3.org/2001/XMLSchema#decimal <‘Rain> <is> <High>
<Intensive’>
```

Two instances are inferred from the triples, which clearly indicate the rainfall values, i.e. 51, and 60.0. These instances are inferred as the instances of high intensive rain.

The Jena rule engine gives the same result as the SWRL rule did, however the difference is the reasoning efficiency and the scalability of Jena over SWRL approach. The Jena inference rule-based system is more scalable in terms of addition of other rules because the rules are not written in an ontology instead defined in a text file and fed into the application. Hence, inference rules serve as a modular unit of knowledge and gives better performance than the SWRL rules.

The third condition that needs to be met to derive the risk of pollution event is to check whether sheep are found in the field. An instance of the output of the sheep node is defined, which is classified by sheep location sensor output. This output has a value in the form of geo coordinates which captures the latitude and longitude of sheep in the field. These coordinates are compared with the geo coordinates of the field. If the sheep coordinates are matched or found within the field coordinates, then it means sheep are in the field. The inference rule in Jena is written as below:

[**rule3SheepInField**:(?s rdf:type enviot:SheepNodeOutput), (?s enviot:isClassifiedBy ?senout),(?senout rdf:type enviot:GroveGPSSheepLocationSensorOutput), (?senout ssn:hasValue ?slval), (?slval rdf:type enviot:GroveGPSSheepLocationValue), (?slval enviot:hasLongitude ?slong), (?slval enviot:hasLatitude ?slat), (?f rdf:type enviot:Field), (?f wgs84_pos:location ?point), (?point rdf:type wgs84_pos:Point), (?point enviot:hasLongitude ?flong), (?point enviot:hasLatitude ?flat), equal(?slong,

?flong), equal(?slat, ?flat) -> print(?slong, ?slat, ‘Sheep’ Coordinates’, ‘found in the Field’s Coordinates’, ?flong, ?flat)]

A sample of the result of the above rule is shown below:

```
'-3.783065'^^http://www.w3.org/2001/XMLSchema#float
'53.202158'^^http://www.w3.org/2001/XMLSchema#float
<‘Sheep’> <Coordinates’> <‘found> <in> <the> <Field’s> <Coordinates’>
'-3.783065'^^http://www.w3.org/2001/XMLSchema#float
'53.20216'^^http://www.w3.org/2001/XMLSchema#float
```

In the above result, the first two lines show the geo coordinates of a particular sheep. The last two lines show the geo coordinates of the field. As the two coordinates match, hence are inferred by the Jena reasoner as the same. Hence, the sheep are found in the field.

Another advantage of this approach is the appropriateness of the inference rules which provides a flexible way of reasoning. Inference rules can easily be modified without affecting the ontology or application. Similarly, new rules can easily be added once new knowledge in the ontology is described.

The fourth rule is about deducing the risk of a pollution event. The first three sub-events in the definition of the pollution event (Figure 5.2) have been defined in the above three inference rules. Now these conditions need to be combined in one inference rule, in addition to a condition about the riparian zone. The inference rule is defined below:

[**rule4RiskOfPollution**:(?s rdf:type enviot:SoilNodeOutput),(?s enviot:isClassifiedBy ?soilsenout), (?soilsenout rdf:type enviot:GroveSoilMoistureSensorOutput), (?soilsenout ssn:hasValue ?gmval), (?gmval rdf:type enviot:GroveSoilMoistureValue), (?gmval enviot:hasSoilMoistureValue ?smval), greaterThan(?smval, 600), (?w rdf:type enviot:WeatherMonitoringDeviceOutput), (?w enviot:isClassifiedBy ?wsenout), (?wsenout rdf:type enviot:CampbellRainfallSensorOutput), (?wsenout ssn:hasValue ?crval), (?crval rdf:type enviot:CampbellRainfallValue), (?crval enviot:hasRainfallValue ?rval), (?hir rdf:type enviot:HighIntensiveRain), (?hir enviot:hasRainfallValue ?hirval) ,ge(?rval, ?hirval), (?sno rdf:type enviot:SheepNodeOutput), (?sno enviot:isClassifiedBy ?sheepsenout),(?sheepsenout rdf:type enviot:GroveGPSSheepLocationSensorOutput), (?sheepsenout ssn:hasValue ?slval), (?slval rdf:type

enviot:GroveGPSSheepLocationValue), (?slval enviot:hasLongitude ?slong), (?slval enviot:hasLatitude ?slat), (?f rdf:type enviot:Field), (?f enviot:hasRiparianZone 'false'^^xsd:boolean), (?f wgs84_pos:location ?point), (?point rdf:type wgs84_pos:Point), (?point enviot:hasLongitude ?flong), (?point enviot:hasLatitude ?flat), equal(?slong, ?flong), equal(?slat, ?flat) -> print(?smval,'Soil is Saturated',?rval, 'Rain is High Intensive',?slong, ?slat, 'Sheep Coordinates match with Fields', ?flong, ?flat, ?f, 'has no riparian zone','ALARM, Risk Of Pollution')]

The result of the above rule is shown below:

```
'625'^^http://www.w3.org/2001/XMLSchema#integer <'Soil> <is> <Saturated'>
'60.0'^^http://www.w3.org/2001/XMLSchema#decimal <'Rain> <is> <High> <Intensive'>
'-3.783065'^^http://www.w3.org/2001/XMLSchema#float
'53.202158'^^http://www.w3.org/2001/XMLSchema#float
<'Sheep> <Coordinates> <match> <with> <Fields>
'-3.783065'^^http://www.w3.org/2001/XMLSchema#float
'53.202158'^^http://www.w3.org/2001/XMLSchema#float
http://www.environmental-iot.com/enviot_ontology/IotSemanticModel#Field1
<'has> <no> <riparian> <zone'>
<'ALARM> <Risk> <Of> <Pollution'>
```

In the above result, the first value (625) is about the soil moisture condition which is greater than 600 and hence the rule is fired saying the soil is saturated. The second line shows the rainfall measurement value (60.0) which falls in the range of high intensive rain defined in the ontology, and hence the condition of high intensive rain is also met. The sheep's coordinates are shown in the fourth and fifth lines, which are the same as that of the field's coordinates in the seventh and eighth lines. Hence, sheep are found in the field. The last condition is about the field's riparian zone which is also met. All the conditions in the antecedent are met and the inference rule is fired inferring the risk of pollution in the catchment by printing the message <'ALARM> <Risk> <Of> <Pollution'>.

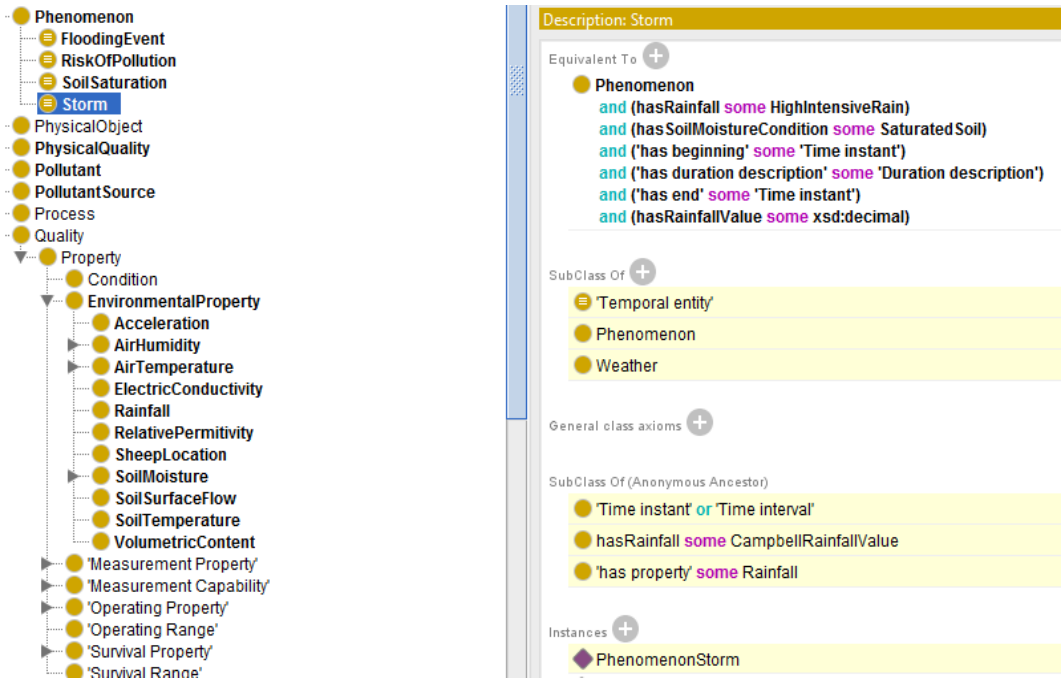


Figure 5.11: Inferring Storm Desmond

### 5.4.2 Evaluating Use-case 2: Geospatial Data Integration and Reasoning

RDF and SPARQL provide support to retrieve data where the relationships are explicitly mentioned. However, sometimes the data is implicitly related to other data, e.g. geospatial data. This leads to a challenge when such data requires integration and reasoning support while retrieving the implicit relationships between disparate datasets [203]. Such data requires indexing and spatial properties and functions to be retrieved. Hence, there is a need to integrate spatial indexing with the inferential power of linking RDF data. To address this challenge, an ontology is required to describe spatial objects supplemented by spatial predicates and functions to retrieve these objects. For this purpose, the OGC GeoSPARQL ontology is adopted and extended in the ontology (see Chapter 4, section 4.4.2(b) and 4.5.6).

In order to perform geospatial data integration and reasoning to fulfil the needs of environmental scientists for the underlying Environmental IoT Infrastructure, a comprehensive set of questions is made. This set of questions is based on the requirements of scientists in the catchment area and contains complex hierarchies and geospatial relations, which can be expressed completely by a geospatial database system. These questions are derived from meetings with environmental scientists and

comprise spatial, temporal and spatio-temporal relations. Some of the questions from each category are given below.

**Spatial Queries:**

- Are soil sensing nodes deployed in the field?
- How many soil sensing nodes are deployed in the field?
- How many soil sensing nodes are deployed in each zone of the field?
- Which soil sensing nodes are deployed in which zone of the field?
- Find soil nutrients in the Hilltop zone.
- Find soil nutrients along with their quantities and measurements units in the Swale region.
- Which geographic zone shows the most likely concentration of Nitrogen?

**Temporal Queries:**

- Find the Storm Desmond start date, end date and rainfall value on those days.
- How long did the Storm Desmond last for?
- What are the rainfall measurements for the month of October?
- Which day recorded the maximum rainfall value during the Storm Desmond?

**Spatio-temporal Queries:**

- Are sheep found in the field during the Storm Desmond?
- How many sheep were found in each zone of the field?
- In which geographic region (zone) the highest number of sheep was recorded during the Storm Desmond?

As mentioned in section 5.2, to store and retrieve RDF triples, GraphDB is used, which is accessed through a Jena application via APIs. Again, the ontology is plugged into the Jena application framework along with the triples. There are 0.2 million triples stored in GraphDB. To see how this approach supports the above complex geospatial queries, the evaluation from each of the spatial, temporal and spatio-temporal queries is carried out in turn.

**Spatial Queries**

**Query-1: Find whether Soil Nodes are deployed in the field.**

In order to check whether sensor nodes are deployed in a given field, the query fetches the geometry of the soil sensing node and compares it with the geometry of the field using the topological function geof:sfIntersects. The hasGeometry property links the node with its geometry using the class geosparql:Point. An individual of the class geosparql:Point takes the geometry form as Point(longitude latitude). The query is written below:

```
ASK WHERE {
        ?node rdf:type enviot:SoilSensingNode;
        geosparql:hasGeometry ?geo.
        ?geo   rdf:type sf:Point;
        geosparql:asWKT ?gwkt.
        ?feature rdf:type enviot:Field;
        geosparql:hasGeometry ?fgeo.
        ?fgeo  geosparql:asWKT ?fwkt.
        FILTER (geof:sfIntersects (?gwkt,?fwkt)) }
```

The output of the query is shown in Figure 5.12.


Figure 5.12: Asking a Question

ASK query gives the output as a Boolean value which is either Yes or No. It can be seen that this approach of retrieving complex geospatial data is not only powerful but also easy to understand and to use. Furthermore, the time taken by this query is shown in the top right corner of the output.

**Query-2: How many Soil Sending Nodes are deployed in each zone of the field, i.e. Hilltop, Swale and Riverbank?**

```
SELECT ?feature   (COUNT (?node) AS ?nodecount) WHERE {
        ?node rdf:type enviot:SoilSensingNode;
        geosparql:hasGeometry ?geo .
        ?geo   rdf:type sf:Point;
        geosparql:asWKT ?gwkt .
        ?feature rdf:type enviot:Field ;
        geosparql:hasGeometry ?fgeo .
        ?fgeo   geosparql:asWKT ?fwkt.
        FILTER (geof:sfIntersects(?gwkt,?fwkt))
        }
    GROUP BY ?feature
    ORDER BY ASC(?nodecount)
```

To remind the reader, the class enviot:Field is defined as the sub-class of geosparql:Feature class in the ontology as described in Chapter 4 (section 4.5.6). As mentioned in Chapter 4 (section 4.4.2), the field is divided into three zones, i.e. Hilltop, Swale, and Riverbank having in total 15 soil sensing nodes. The above spatial query retrieves all this knowledge modelled in the ontology. The result is shown below in Figure 5.13.

Filter query results — Showing results from 1 to 3 of 3. Query took 0.147 s.

| | feature | nodecount |
|---|---|---|
| 1 | enviot:A_Hilltop | "3"^^xsd:integer |
| 2 | enviot:A_Riverbank | "5"^^xsd:integer |
| 3 | enviot:A_Swale | "7"^^xsd:integer |

Figure 5.13: Illustration of Soil Sensing Nodes in Each Zone of the Field

The above result endorses the sketch map of the sensor nodes deployed in the catchment, as shown in Chapter 4 (see Figure 4.6). An important point to mention here is the use of the filter function geof:sfIntersect in the above query, which can be replaced by geof:sfWithin. The filter function geof:sfWithin in the above query in fact makes more sense instead; however geof:sfIntersect is used owing to the issue of lack of geo coordinates of the field.

**Query-3: Find soil nutrients in the Hilltop region.**

```
SELECT ?soilnutrient ?geo ?feature WHERE
  {
  ?soilnode rdf:type enviot:Soil ;
  enviot:hasSample  ?handsample.
  ?handsample  enviot:hasVariable ?soilnutrient;
  geosparql:hasGeometry ?ngeo.
  ?ngeo geosparql:asWKT ?geo .
  ?feature rdf:type  enviot:Hilltop;
  geosparql:hasGeometry ?fgeo.
  ?fgeo geosparql:asWKT ?fwkt.
 FILTER(geof:sfIntersects(?geo, ?fwkt))
  }
 LIMIT 09
```

One of the potential strengths of the proposed approach is the rich knowledge modelling capability of the ontology about the domain. In order to test the quality of soil sensor measurements, soil scientists also collect the hand sample about soil nutrients in the catchment and bring them back to the lab for analysis. This knowledge is described in the ontology using the classes enviot:SoilHandSample and enviot:SoilVariable. The result shows all soil nutrients along with their geometries and the region (Hilltop) where they are collected (Figure 5.14).

Filter query results — Showing results from 1 to 9 of 9. Query took 0.133 s.

| | soilnutrient | geo | feature |
|---|---|---|---|
| 1 | enviot:Ammonium | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 2 | enviot:Calcium | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 3 | enviot:ElecConductivity | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 4 | enviot:MoistureContent | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 5 | enviot:Nitrate | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 6 | enviot:Phosphate | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 7 | enviot:Potassium | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 8 | enviot:Sodium | POINT(-3.781111 53.202777) | enviot:A_Hilltop |
| 9 | enviot:SoilpH | POINT(-3.781111 53.202777) | enviot:A_Hilltop |

Figure 5.14: Soil Nutrients in the Hilltop Zone

**Query-04: Find soil nutrients in the Swale region along with their quantities and measurement units?**

```
SELECT  DISTINCT ?feature ?soilnutrient ?quantity ?unit WHERE
{
  ?soilnode rdf:type  enviot:Soil ;
  enviot:hasSample ?handsample.
  ?handsample  enviot:hasVariable ?soilnutrient;
  geosparql:hasGeometry ?ngeo.
  ?ngeo   geosparql:asWKT ?nwkt.
  ?soilnutrient   enviot:hasQuantityValue ?quantity;
  enviot:hasQuantityUnitOfMeasurement ?unit.
  ?feature  rdf:type  enviot:Swale;
  geosparql:hasGeometry ?fgeo.
  ?fgeo   geosparql:asWKT ?fwkt.
  FILTER(geof:sfIntersects(?nwkt, ?fwkt))
 }
 ORDER BY ?soilnutrient
```

A sample of the result of the above query is shown in Figure 5.15.

| | feature | soilnutrient | quantity | unit |
|---|---|---|---|---|
| 1 | enviot:A-Swale | enviot:Calcium | "8.287E1"^^xsd:double | "enviot:parts-per-million" |
| 2 | enviot:A-Swale | enviot:ElecConductivity | "1.4E1"^^xsd:double | "enviot:parts-per-million" |
| 3 | enviot:A-Swale | enviot:ElecConductivity | "1.4E1"^^xsd:double | "enviot:parts-per-million" |

Figure 5.15: Soil Nutrients in Swale along with Quantities and Metric Units

Another important potential quality feature of this approach is the assignment of accurate and automated associated metric units alongside their quantities. Not only sensor measurements are semantically enriched but the associated metric units are also enriched with the ontology. This feature is discussed in more detail in use-case 3.

**Query-05: Which geographic feature shows the most likely concentration of Nitrogen?**

```
SELECT  DISTINCT ?feature (MAX(?nitrate) AS ?MaxConcentration) WHERE {
?soilnode rdf:type enviot:Soil ;
enviot:hasSample ?handsample.
?handsample enviot:hasVariable ?soilnutrient;
geosparql:hasGeometry ?ngeo.
?ngeo geosparql:asWKT ?nwkt.
enviot:Nitrate enviot:hasQuantityValue ?nitrate.
?feature rdf:type enviot:Field;
geosparql:hasGeometry ?fgeo.
?fgeo geosparql:asWKT ?fwkt.
FILTER(geof:sfIntersects(?nwkt, ?fwkt)) }
GROUP BY ?feature
```

The result of the above query is shown below in Figure 5.16:


Figure 5.16: The Most Likely Concentration of Nitrogen in the Field

One of the great potential features of this approach is combining the strength of both the ontology reasoning support with GeoSPARQL topological relationships functions. Using GeoSPARQL filter functions, e.g. geof:sfIntersect, geof:sfWithin etc. for topological comparisons between different geometries, makes the cumbersome task of complex data integration and geospatial reasoning really easy.

The next couple of examples show temporal queries.

## Temporal Queries

### Query-06: Storm Desmond Start and End Date and Duration

```
SELECT ?storm ?startdate ?enddate ?durationdays ?durationhours WHERE {
?storm rdf:type enviot:Storm ;
time:hasBeginning ?begin;
time:hasEnd ?end;
time:hasDurationDescription ?duration.
?begin time:inXSDDateTime ?startdate.
?end time:inXSDDateTime ?enddate.
?duration time:days ?durationdays;
time:hours ?durationhours. }
```

In order to show the temporal characteristics of the domain, a storm event, i.e. Storm Desmond was modelled in the ontology and was successfully inferred in the previous section. To conceptualise the information about the storm event, the W3C Time ontology is used, which provides temporal properties to describe such events. The result of the query is shown in Figure 5.17.

Filter query results — Showing results from 1 to 1 of 1. Query took 0.186 s.

| | storm | startdate | enddate | durationdays | durationhours |
|---|---|---|---|---|---|
| 1 | enviot:A_StormDesmond | "2015-12-04"^^xsd:date | "2015-12-06"^^xsd:date | 2 | "48"^^xsd:float |

Figure 5.17: Retrieval of Storm Desmond

The result provides information about Storm Desmond that started on 4th December 2015 and ended on 6th December that year, lasting for two days (48 hours). From the result, it is confirmed that the structure of the ontology not only provides support to reason over events but also to retrieve temporal information about those events.

### Query-07: What Rainfall Measurements were recorded on start and end dates of Storm Desmond?

```
SELECT ?storm ?startdate ?enddate ?rainfallstartvalue ?rainfallendvalue WHERE {
?storm rdf:type enviot:Storm;
time:hasBeginning ?begin;
time:hasEnd ?end;
```

```
time:hasDurationDescription ?duration.
?begin time:inXSDDateTime ?startdate;
enviot:hasRainfallValue ?rainfallstartvalue.
?end time:inXSDDateTime ?enddate;
enviot:hasRainfallValue ?rainfallendvalue. }
```

The result of the above query is shown in Figure 5.18. The result shows the rainfall measurements collected by the sensors during the storm's dates.

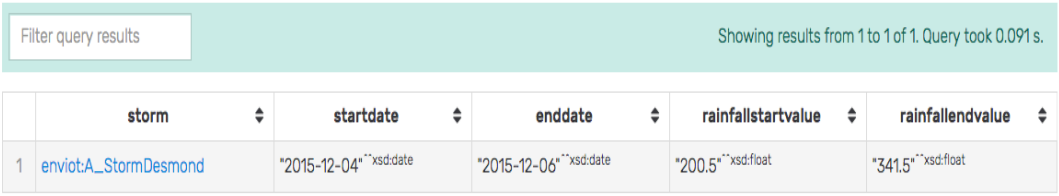


Figure 5.18: Rainfall Measurements on Storm Desmond Dates

It can be seen from the results that the knowledge represented in the ontology provides strong support to not only spatial (successful information retrieval of first five queries) but also temporal (last two temporal queries) events. The approach also provides querying support to retrieve temporal properties about these events on a more fine-grained level including hours, minutes and seconds. This allows that other complex extreme events can also be formalised in the ontology and can successfully be retrieved by using this approach.

So far, the examples have shown the spatial and temporal knowledge retrieval from the system about the domain. The next set of examples show some more complex queries, which would combine both the spatial and temporal characteristics of knowledge in one query to perform spatio-temporal data integration and reasoning about the events occurring in the catchment area.

**Spatio-temporal Queries**

**Query-08: Are sheep found in the field during the Storm Desmond?**

```
ASK WHERE {
        ?sheep rdf:type enviot:SheepNodeOutput ;
        time:inDateTime ?instant;
        geosparql:hasGeometry ?sgeo.
        ?sgeo  geosparql:asWKT          ?swkt.
```

```
        ?instant time:inXSDDateTime ?sdate.
        ?storm rdf:type enviot:Storm;
        time:hasBeginning      ?begin.
        ?begintime:inXSDDateTime ?stormdate.
        ?feature rdf:type enviot:Field;
        geosparql:hasGeometry ?fgeo.
        ?fgeo  geosparql:asWKT         ?fwkt.
    BIND(xsd:date(concat(str(year(?sdate)),"-", str(month(?sdate)),"-",
         str(day(?sdate)))) AS ?sheepdate)
     FILTER (geof:sfIntersects(?swkt, ?fwkt))
     FILTER (?sheepdate = ?stormdate) }
```

The above query combines all three dimensions of knowledge representation described in Chapter 4 (see section 4.4.2), i.e. thematic (sheep), spatial (field) and temporal (storm's instant). Furthermore, in the query a new function Bind () is also introduced, which takes different parts of the date, which is in the form of string, and converts those parts separately to correct date format and then concatenates all these constituent parts. The reason for doing this conversion is that the scripts converted the data from JSON to JSON-LD format but due to some reason the date was expressed in the string format. The result of the query is shown in Figure 5.19.


Figure 5.19: Sheep Found in the Field during the Storm Desmond

The above successful retrieval of the complex event, in addition to reasoning performed in use-case 1, further confirms that the knowledge modelled in the ontology is functional. The ease of use of querying support provided by the approach is another huge advantage.

**Query-09: How many Sheep were found in each zone of the field (i.e. Hilltop, Swale, and Riverbank) during Storm Desmond?**

```
SELECT DISTINCT ?feature (COUNT(?sheep) AS ?sheepcount) WHERE {
        ?sheep rdf:type enviot:SheepNodeOutput ;
        time:inDateTime ?instant;
        geosparql:hasGeometry ?sgeo.
        ?sgeo  geosparql:asWKT ?swkt.
        ?instant time:inXSDDateTime ?sdate.
        ?storm rdf:type enviot:Storm;
        time:hasBeginning ?begin.
        ?begintime:inXSDDateTime ?stormdate.
        ?feature rdf:type enviot:Field;
        geosparql:hasGeometry ?fgeo.
        ?fgeo  geosparql:asWKT ?fwkt.
BIND(xsd:date(concat(str(year(?sdate)),"-", str(month(?sdate)),"-",
str(day(?sdate)))) AS ?sheepdate)
FILTER (geof:sfIntersects(?swkt, ?fwkt))
FILTER (?sheepdate = ?stormdate)}
GROUP BY ?feature ?sheep ?sdate
ORDER BY ASC (?sheepcount)
```

So far, this is the most highly complex query which performs data integration and reasoning over thematic, spatial and temporal data and involves almost all predicates of the SPARQL query. The result of the query is shown in Figure 5.20.

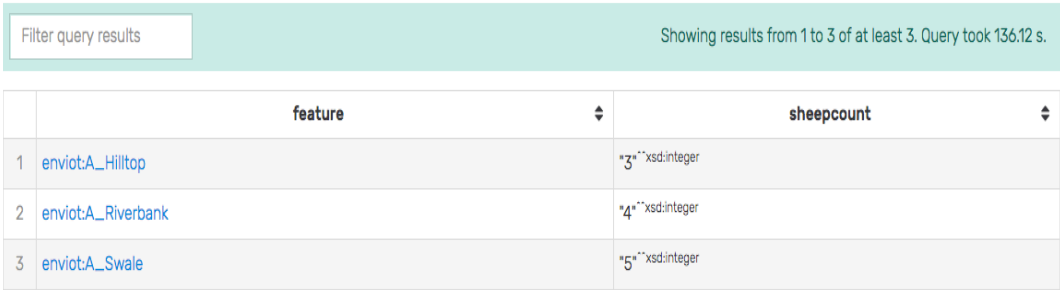


Figure 5.20: Sheep in the Field during Storm Desmond

The result accurately finds the number of sheep in each zone of the field, i.e. Hilltop, Riverbank and Swale, during Storm Desmond. The query further validates the consistent structure of concepts and their relationships and the precise modelling of knowledge about the domain in the ontology. However, the query takes relatively more time to perform this complex spatio-temporal reasoning. There are a few reasons for this inefficiency: i) as mentioned above, the query involves all three

dimensions of knowledge representation; ii) the date data is in string format due to which the Bind function () introduces additional complexity; iii) almost all predicates of SPARQL query are used; iv) the query involves the topological comparison of geospatial function (geof:sfIntersects). The efficiency can be improved to some extent by using the correct date format, however further algorithmic techniques are required to optimise the performance of information retrieval, which is beyond the scope of this work.

**Query-10: Which geographic region (i.e. Swale, Hilltop and Riverbank) observes the maximum number of sheep in the field during the Storm Desmond?**

```
SELECT DISTINCT ?feature (COUNT(?sheep) AS ?sheepcount) WHERE {
        ?sheep rdf:type enviot:SheepNodeOutput ;
        time:inDateTime ?instant;
        geosparql:hasGeometry ?sgeo.
        ?sgeo  geosparql:asWKT ?swkt.
        ?instant time:inXSDDateTime ?sdate.
        ?feature rdf:type enviot:Field;
        geosparql:hasGeometry ?fgeo.
        ?fgeo  geosparql:asWKT ?fwkt.
        ?storm rdf:type  enviot:Storm;
        time:hasBeginning ?begin.
        ?begintime:inXSDDateTime ?stormdate.
  BIND (xsd:date(concat(str(year(?sdate)),"-", str(month(?sdate)),"-",
  str(day(?sdate)))) AS ?sheepdate)
  FILTER(geof:sfIntersects(?swkt, ?fwkt))
  FILTER (?sheepdate = ?stormdate) }
  GROUP BY ?feature ?sheep ?sdate
  ORDER BY DESC(?sheepcount)
  LIMIT 1
```

It can be noted that the query does not involve the Max () function. This is because this function is not fully supported in some cases particularly in this case.

The result of the above query is shown in Figure 5.21.

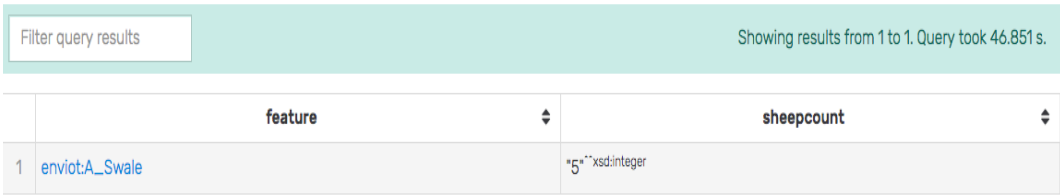


Figure 5.21: Maximum Sheep Count in the Region during Desmond Storm

The functional correctness of the ontology can be confirmed from the result of the above query in Figure 5.21 that can further be confirmed from the result of the previous query shown in Figure 5.20.

### 5.4.3 Evaluating Use-case 3: Interoperable Metric Units

In order to evaluate the ontology to provide unambiguous interoperable metric units, initially the MUO/UCUM is used and extended for the metric units of those physical qualities that are not described in the MUO ontology. After importing the MUO/UCUM ontology, several SWRL rules are written which would unambiguously translate one unit to another. Initially, the SWRL rules are written to test the conversion process, followed by then the corresponding Jena inference rules. The first SWRL rule is written to translate the measurement value of air temperature in degree Celsius to degree Fahrenheit. In the rule, 'x' is an instance of the class which represents an air temperature value in degree Celsius. The value is then multiplied by 1.8 and is stored in another variable 'product'. The resultant value is then added with 32 and is stored in a variable 'sum'.

**SWRL Rule1**: Converting Degree Celsius to Degree Fahrenheit

GroveAirTemperatureValue(?x)^hasTemperatureInCelsius(?x,?tempval)^hasUnitOf Measurement(?x, DegreeCelsius)^swrlb:multiply(?product, ?tempval, 1.8) ^ swrlb:add(?sum, ?product, 32) -> enviot:hasTemperatureInFahrenheit(?x, ?sum) ^ hasUnitOfMeasurement(?x, DegreeFahrenheit)

In order to perform the above unit conversion, the Pellet reasoner is selected and is computed to reason over the ontology. It is observed that the reasoning efficiency is very low due to a large number of instances/members in the MUO/UCUM ontology. Hence, keeping in account the ontology design guidelines (Chapter 4, section 4.2), it is decided that instead of using and extending MUO/UCUM, a minimal lightweight

ontology should be designed for metric units. Thus, MUO/UCUM ontology is dropped and a new lightweight ontology is developed (see Chapter 4, section 4.5.8). In order to reason over the new metric units ontology for metric unit conversion interoperability, the reasoner is started again. The result of the reasoning is shown Figure 5.22.

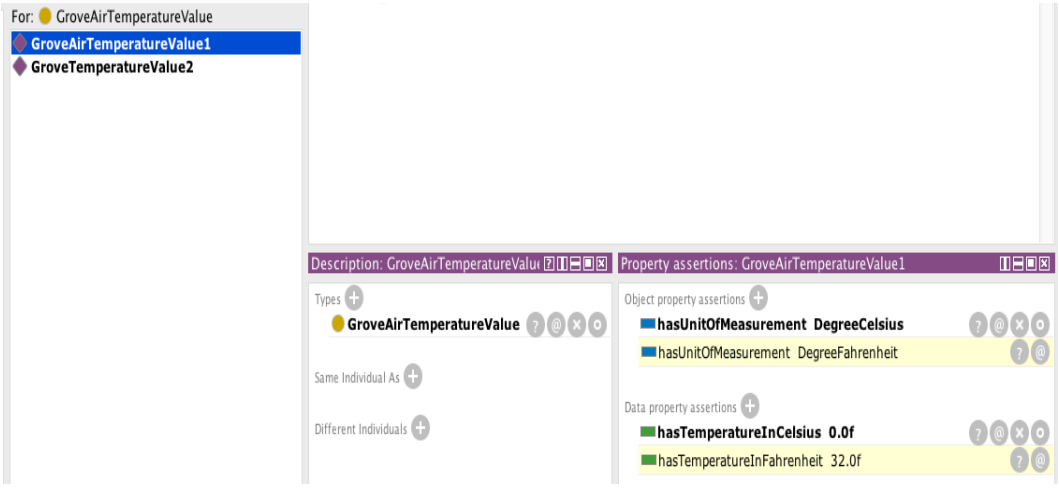


Figure 5.22: Conversion of Degree Celsius to Degree Fahrenheit

As can be seen from the result in Figure 5.22. above, the actual value of the instance GroveAirTemperatureValue in the ontology is stored in degree Celsius, which is 0.0 (f represents that the value is float) degree centigrade. After the reasoner is run, it can be seen (in yellow) that both the equivalent unit and quantity value are classified and converted as 'DegreeFahrenheit' and 32.0f respectively.

Similarly, the corresponding rule from Fahrenheit to degree Celsius is defined as:

**SWRL Rule 2**: Converting degree Fahrenheit to degree Celsius

enviot:GroveAirTemperatureValue(?x) ^ enviot:hasUnitOfMeasurement(?x, enviot:DegreeFahrenheit) ^ enviot:hasTemperatureInFahrenheit(?x, ?tempval) ^ swrlb:divide(?div, ?sub, 1.8) ^ swrlb:subtract(?sub, ?tempval, 32)
-> enviot:hasUnitOfMeasurement(?x, enviot:DegreeCelsius) ^ enviot:hasTemperatureInCelsius(?x, ?div)

The result is show in Figure 5.23. An individual named GroveTemperatureValue2 is defined having a temperature measurement value 212.0 in degree Fahrenheit, which is

accurately converted to 100.0 degree Centigrade along with the metric unit DegreeCelsius. Both these inferred facts are shaded in yellow.

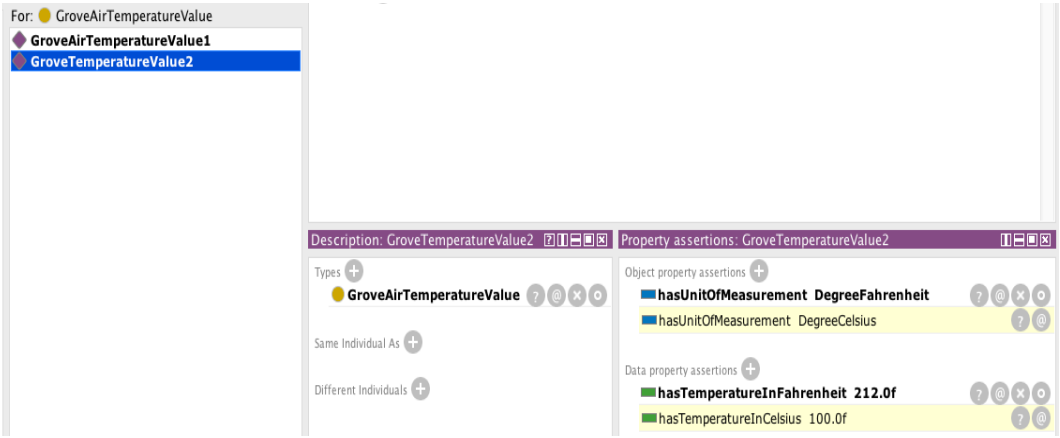


Figure 5.23: Conversion of Degree Fahrenheit to Degree Celsius

One of the strengths of this approach is the semantic annotation of measurement quantities in a simple, efficient and unambiguous way. The measurement quantities about different environmental physical qualities are collected by sensors without associated metric units. These measurements are then assigned their associated metric units with the help of the ontology. Environmental scientists are not concerned anymore regarding the unit conversion interoperability. They can use any standard SI units they want. This assignment is modelled in the ontology as shown in Figure 5.24.

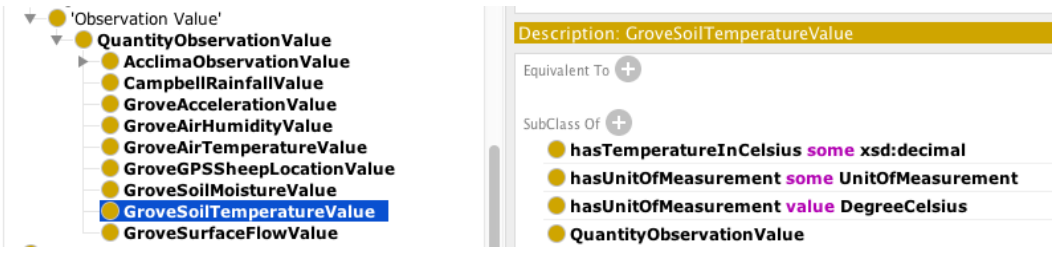


Figure 5.24: Description of the class enviot:GroveSoilTemperatureValue

The alternative flexible and efficient way to metric unit conversion interoperability is using Jena rules. The above SWRL rule 1 is defined in Jena as under:

[**JenaRule1**: (?t rdf:type units:GroveAirTemperatureValue), (?t enviot:hasTemperatureInCelsius ?tempval), (?t enviot:hasUnitOfMeasurement enviot:DegreeCelsius), product(?tempval, 1.8, ?prod), sum(?prod, 32, ?add) -> (?t enviot:hasTemperatureInFahrenheit ?add), (?t enviot:hasUnitOfMeasurement units:DegreeFahrenheit)]

The result of the Jena rule that converts both metric units and its quantity value from degree Celsius to degree Fahrenheit is shown below:

```
<http://www.environmental-
iot.com/enviot_ontology/IotSemanticModel#GroveAirTemperatureValue1>
<http://www.environmental-
iot.com/enviot_ontology/IotSemanticModel##hasTemperatureInFahrenheit>
"32.0"^^<http://www.w3.org/2001/XMLSchema#double>;
<http://www.environmental-
iot.com/enviot_ontology/IotSemanticModel#hasUnitOfMeasurement>
<http://www.environmental-
iot.com/enviot_ontology/IotSemanticModel#DegreeFahrenheit>.
```

The quantity value of the individual GroveAirTemperatureValue1 is 0 degree Centigrade which is successfully converted to 32.0 degree Fahrenheit. Similarly, like rule 1, a set of Jena inference rules is written which performs unambiguous metric unit conversion for all the physically qualities used in this work.

## 5.5 Overall Evaluation

This section presents the overall evaluation of the approach across all use-cases with regard to the evaluation criteria described in section 5.3.

Regarding the **functional qualitative parameter**, the ontology is evaluated to check whether it accomplishes all the tasks for which it is designed. In both use-case 1 and use-case 3, the functional parameter is evaluated by looking at its inference ability in the tasks of deriving a pollution event and unambiguous metric units conversion respectively. The definitions of all events were defined by environmental scientists, and were modelled accordingly in the ontology by the author. Two types of rules were defined: the SWRL rules and the Jena inference rules. The conditions used in the antecedent of the rules were matched against the triples in the knowledgebase and the actions in the consequent were fired and inferred the required new facts. The SWRL rules were processed by the in-built Pellet reasoner in the Protégé ontology editor and the Jena inference rules were processed by the general-purpose rule-based engine in

the Jena application framework. The results showed that all the events were inferred successfully. The results further showed that all metric units were converted unambiguously and correctly to other metric system along with their quantities. Regarding the functional parameter of use-case 2, all three dimensions of knowledge representation of the target domain were defined in the queries to fulfil the information needs of environmental scientists. SPARQL queries were used with the geospatial predicates and functions of the GeoSPARQL ontology. The results clearly showed that the queries retrieved the knowledge and answered complex questions of environmental scientists. The spatio-temporal data was effectively integrated from a wide variety of sources and the geospatial reasoning was performed successfully. Minor deficiencies were found during all the three use-cases regarding the formation of the rules and queries. However, the deficiencies were addressed during the iterative process as part of the ontology development. The evaluation showed that the ontology fulfilled the functional purpose.

**Expressiveness**, in the context of this thesis, is defined as how well the ontology performs the full range of tasks when used within the target domain. The expressive parameter of the ontology is evaluated by looking at its natural support it provides to environmental scientists, e.g. in terms of writing an inference rule or a query. The inference rules followed the 'If Then Else' form of structured English, and hence were really easy to understand and write. Most of the scientists knew already about such rules. To keep the approach simple, the Jena inference rules were written in a separate text file instead of writing them in Jena application. This served two main purposes: i) easy addition of other rules and maintenance; ii) making the rules easy for scientists to understand and write their own rules with little technical knowledge. From the evaluation, it can be seen that the inference rules were natural to expressing the knowledge of events in the domain. Most scientists would understand both the ontology and inference rules but, still they would need some time and technical knowledge to understand both. Regarding SPARQL queries, the GraphDB triplestore was used for storing and retrieving triples. Although GraphDB was accessed from the Jena application through APIs, however because of the Workbench support of GraphDB, its user-friendly GUI interface made it natural for scientists to understand the queries. The syntax of SPARQL query is very similar to SQL, which further simplified the querying component and information retrieval for scientists to write

their own queries with little SQL and RDF knowledge. Furthermore, due to the simple structure of RDF triples in the form of subject, predicate and object, it would be quite natural for most of the scientists to understand the queries and RDF triples.

To take the evaluation further, how well the ontology serves the purpose in terms of supporting other pertinent events in the catchment, the ontology was evaluated for deriving an extreme event, i.e. Storm Desmond that occurred from 4$^{th}$ to 6$^{th}$ December 2015. To check whether the system deduces a given storm event, a sub-class called enviot:Storm of the enviot:Phenomenon class was created and was made as a defined class. The purpose of this modelling was to see whether the ontology infers a storm event (e.g. Storm Desmond). A SWRL rule was defined in the ontology to infer this event. Similar to the risk of pollution event, the storm event was also dependent on some other events, e.g. high intensive rainfall, soil saturation, and some temporal characteristics. These events were described in the ontology with some additional temporal knowledge. An instance, named PhenomenonStorm, of the class enviot:Phenomenon was created, which satisfied the definition of the defined class enviot:Storm. To deduce the storm event, the rule was defined as:

enviot:Phenomenon(?p)^enviot:hasRainfall(?p,enviot:HighIntensiveRain2)^
enviot:hasSoilMoistureCondition(?p,enviot:SaturatedSoilStorm)^
time:hasBeginning(?p,enviot:StormDesmondStart) ^ time:hasDurationDescription(?p, enviot:StormDesmondDurationDescription)^time:hasEnd(?p,enviot:StormDesmondEnd) –> enviot:Storm(?p)

After running the reasoner over the ontology, the result illustrated that the instance (PhenomenonStorm) was classified successfully under the class enviot:Storm descriptions, as shown in Figure 5.25.

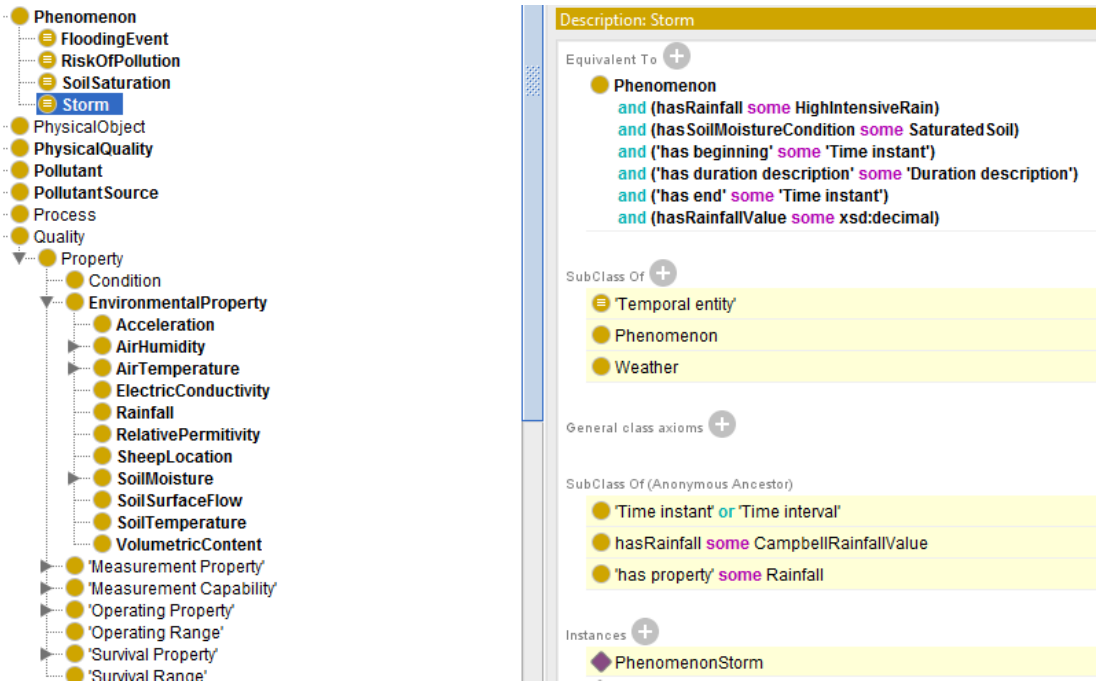


Figure 5.25: Inferring Storm Desmond

From the evaluation of the results, it is demonstrated that the approach has some **strengths**, however the major one is its ability of modularisation and appropriateness for the target domain. Modules can easily be added in the ontology to conceptualise other extreme events, for instance, storm. This factor has already evaluated in terms of adding a module describe a storm event and then inferring one particular instance of a storm, i.e. the Storm Desmond event. Based on this evaluation, the author postulates that the ontology can be adopted in other sub-domains of environmental science with minimal addition of domain knowledge in the ontology. For instance, it can be extended in the hydrology domain for deriving the risk of water pollution.

To summarise, from the evaluation it can be seen that the ontology largely satisfies the primary data needs of environmental scientists for the underlying Environmental IoT Infrastructure. This has been observed in evaluating the results of all three use-cases. However, the approached suffered from some limitations that are described below.

The major **limitation** of the approach is the efficiency of the system when it performs particularly spatio-temporal reasoning on SPARQL queries over large data. The time of geospatial reasoning increases as the size of data increases. Though in this work, a full quantitative evaluation of geospatial reasoning was not performed, in general it is

clear that the efficiency needs to be improved. How scalable will the system be when huge heterogeneous data is integrated involving both many spatial and temporal comparisons? There is a need to further improve the proposed approach for spatio-temporal data integration and reasoning particularly for huge data.

The evaluation also discovered some other limitations of the approach regarding use-case 2. One of the limitations of this approach is that it does not perform qualitative geospatial reasoning. In such reasoning systems, the RCC (Region Connection Calculus) topological inference is enabled for features having unknown geometries. Another limitation is the low efficiency of the approach that arises in the information retrieval of only highly complex spatio-temporal queries. Further research is required to address the above said limitations, however, this lies outside the scope of this thesis and is left as a future work.

Regarding use-case 3, the approach works well for providing an unambiguous exchange of quantities alongside their associated units. However, the main limitation of this approach is that it provides unit conversion interoperability only for those physical qualities that have a standard conversion formula, e.g. SI units. The approach lacks support for dealing with more complex phenomena where the mapping between two representations are not well defined. For example, from one of the interviews with a soil scientist, it was discovered that representing measurements of physical qualities using two different models in two sub-disciplines of environmental science (biogeochemistry and plant ecology) encountered issues. In this specific case, simple formula conversion would not work and this instead requires some sophisticated transfer functions to be developed which is lacking in this approach. For the said complicated phenomena, there is a mismatch between the real needs of the scientists and what the approach has been able to show so far. So, the unit conversion interoperability which is a bigger point, is partially solved and can be further improved as a future work to address such complex unit conversion phenomena in environmental science.

## 5.6 Overall Analysis and Lessons Learned

This work provides a wealth of experience in the potential of Semantic Web technologies applied to understand the complex and heterogeneous data in

environmental science. Key lessons were learned from the overall approach, which are described below.

1. The potential of Semantic Web technologies in underpinning environmental science is significant and the approach can provide a much richer and interoperable representation of complex and heterogeneous data. Through ontologies, RDF, linked data and SPARQL, environmental data can be combined and interlinked with external data sources, leading to a data-rich linked open data cloud for environmental science.

2. The combination of Semantic Web and IoT technologies is overarching and natural in underpinning environmental science – with IoT providing a rich set of streaming data covering thematic, spatial and temporal dimensions of the ecosystems, and Semantic Web technologies offering significant capacity in terms of making sense of the rich volumes and variety of data in all its complexity.

3. Designing and developing ontologies for an interdisciplinary and integrative domain like environmental science is very challenging owing to the issue of agreeing on the consistent vocabulary of all interconnected sub-disciplines. Hence, the ontology in environmental science should make minimal ontological commitment, i.e. the ontology should make as few claims as possible about the disciplines being modelled, allowing the communities committed to the ontology development to extend and specialise the ontology as required [62]. The communities should agree on the usage of vocabulary that is consistent (but not complete) in terms of the concepts or theory described by the ontology.

4. Data integration and reasoning have been performed over a small number of triples (0.2 millions) which gives some good results in terms of data integration and inference (takes time in seconds both in integration and reasoning when the query is simple, and a couple of minutes when the query is extremely complex involving thematic, spatial and temporal matching patterns). Thus, increasing the number of triples and filters (in a query) might increase data integration and inference time. Hence, efficiency might be an issue over a huge number of triples (say triples in millions).

5. Data integration and reasoning have been performed taking into account only qualitative parameters because this is required by the scientists (collaborators) in the

context of Environmental IoT Project. However, to check data integration and reasoning efficiency, scalability and reasoning complexity of the approach over millions of triples for the IoT infrastructure may also be technically important.

6. Another important lesson that has been learned regarding reusability of the ontology is that there is a trade-off between usability and reusability of the ontology designed especially for a particular application. The more you describe concepts of a particular domain in an ontology design, the more the ontology becomes specialised and hence less reusable for other applications. In the ontology developed in this thesis for the Environmental IoT Project, some of the concepts (or modules), for instance, sensor modules, device modules, temporal modules, spatial modules and metric unit modules, can be reused in other sub-disciplines of environmental science. However, those modules that are designed specifically for the target application, for instance, the 'phenomenon module' to find the risk of pollution, cannot be reused in other disciplines. Hence, it is hard to achieve both designing an ontology for a specific use or application in a particular domain while preserving reusability of that ontology in other domains simultaneously [204]. This is one of the limitations of an application ontology.

7. SWRL rules provide an additional expressive power to ontological reasoning. However, the more you add SWRL rules, the more it affects reasoning efficiency. Furthermore, SWRL cannot access external data sources and hence the data has to be brought into the ontology. On the other hand, Jena rules are more flexible and powerful in terms of reasoning efficiency and do not add any extra complexity to the ontology.

8. To discover interdependencies between disparate datasets representing different environmental facets, more instrumentation is required in the natural environment to capture rich and ubiquitous streaming data of a vast variety of environmental variables at different geographical locations and at different scale. Due to restricted funding and resources in the Environmental IoT project, there was limitation on the number of environmental variables that could be captured. A subsequent large-scale evaluation would be very interesting.

9. Metric units in any scientific domain are extremely important and quantitative measurements would be incomplete without their associated metric units. To provide semantic interoperability and data integration, the metric units ontology can play a key role. During the ontology design phase, a key decision was made whether to develop a minimal lightweight units ontology from scratch or reuse and extend existing complex unit ontology that are designed for covering a lot of domains. Evaluating the use-case about metric units, it was found that existing unit ontologies have a large number of individuals that lead to inefficiencies in reasoning and computations. On the other hand, a minimal lightweight unit ontology designed for one particular domain performs better in terms of efficiency.

10. When compared to other disciplines like life sciences and bioinformatics, the uptake of Semantic Web technologies, particularly ontologies and linked data in underpinning environmental science, is still low. More research is required to further explore these technologies in combination with other techniques including web services to offer standardised interfaces and machine learning to form intelligent decision support systems in order to make sense of this complex and heterogeneous environmental data.

11. Last but not least, developing a diverse set of skills is required for working in a collaborative multi-disciplinary environment particularly when designing ontology in environmental science. Input of domain experts is really important in the design phase when an application ontology is aimed at accomplishing a particular task especially specifying real-world scenarios, for example, modelling extreme events. Communication and regular contacts with domain experts are crucial to get insights into the domain and break the inevitable language barriers.

## 5.7 Conclusion

This chapter has presented an evaluation of the proposed overall framework comprising the ontology and the Jena application. The experimental evaluation through three real-world use-cases has shown that the framework achieves the main objectives of the thesis, i.e. using Semantic Web technologies to discover interdependencies between disparate datasets, to perform geospatial data integration and reasoning, and to provide interoperability. The work has demonstrated that the

approach is able to identify causal-like relations between different events in the catchment (here inferencing the risk of a pollution event in the catchment), to integrate spatio-temporal data addressing complex queries of environmental scientists and to provide unambiguous interoperable metric units. This evaluation also validates the overarching role of the ontologies in all these three use-case scenarios. Furthermore, the role of the rule-based inference techniques in the application framework is found to be important. In summary, the overall approach described in the experimental evaluation is able to address the three research challenges and that the objectives of the research have been met to a great extent.

# 6 Conclusion

## 6.1 Introduction

This thesis has investigated Semantic Web technologies for IoT/streaming data in underpinning environmental science. More precisely, an ontology was developed and used along with associated Semantic Web techniques to discover the interdependencies between disparate but interlinked datasets, to perform data integration and geospatial reasoning and to provide interoperable unambiguous metric units as an example of dealing with heterogeneity.

This chapter concludes the research by providing a summary of the narrative within the thesis, highlighting the major contributions of the research, and discussing potential areas of future work.

## 6.2 Thesis Summary

**Chapter 1** introduced the thesis by presenting the context of the research and its relevance to the area of Semantic Web technologies. The chapter described the key objectives and the overarching research questions that drove this research. Furthermore, the chapter also discussed briefly the research methodology and finally concluded with the presentation of the thesis's outline.

**Chapter 2** provided a background overview of Semantic Web technologies and explored the state of the art in the use of such technologies and techniques in the

context of eScience. The chapter provided a more in-depth assessment of related work that lies at the intersection of three key areas including the Semantic Web, IoT/streaming data, and environmental science. Finally, the chapter concluded with the argument that there is limited research at the intersection of these three areas and hence further research is required in terms of meeting the particular data needs of environmental science.

**Chapter 3** examined the unique characteristics of environmental science in the context of environmental data, through semi-structured in-depth interviews. The chapter aimed particularly at exploring and collecting qualitative data covering different aspects including: the role of data and practices, data trends, interdependencies between disparate but interlinked datasets, and technological opportunities and barriers in environmental science. The chapter provided the analysis of the qualitative data using a Grounded Theory methodology and concluded with three key findings that then shape the next phase of the research, namely the interdependencies between disparate datasets, geospatial data integration and reasoning, and interoperability.

**Chapter 4** introduced the ontological framework for the environmental IoT data. The chapter provided an overall design of the ontology as well as the integration of other ontologies imported and extended in this work. The chapter also described various key design criteria of the ontology including reusing existing standards, modularity and extensibility, expressiveness and reasoning support and aiming for a lightweight design. Finally, the chapter concluded with the argument that the ontology in environmental science should aim for more lightweight but extensible model that communities can agree with and which can be extended over time as concepts are deemed missing.

**Chapter 5** provided an evaluation of the work through three different real-world use-cases, derived from the analysis of the semi-structured interviews and meetings with environmental scientists. This evaluation was carried out to demonstrate the applicability, strengths and limitations of the ontology and the overall approach in the target discipline(s) of environmental science. The experimental evaluation through three real-world use-cases showed that the approach achieves the main objectives of the thesis, i.e. using Semantic Web technologies to discover interdependencies

between disparate datasets, to perform geospatial data integration and reasoning, and to provide unambiguous metric units interoperability. The evaluation also validated the overarching role of the ontologies in all these three use-case scenarios.

## 6.3 Contributions of the Thesis

The main goal of this research has been to examine the potential role of Semantic Web technologies and their applicability in supporting a deeper understanding of the natural environment as derived from a plethora of sources of environmental data. The thesis has adopted a mixed methods approach involving a literature review, substantive semi-structured interviews and the development of an experimental ontology with environmental streaming data. This has led to the following contributions.

### 6.3.1 Characteristics of Environmental Data

This thesis provides some key insights into the nature of environmental data and the particular challenges associated with this area of science. In particular, the thesis has identified:

- The importance of the concept of the long tail of science as it applies to environmental data. In environmental science, there exists a large number of small, heterogeneous and potentially complex datasets that are usually collected by individual scientists, small laboratories and/or projects. When combined together, they form a big portion of the data spectrum.
- Five key challenges associated with environmental data have been highlighted that make this area quite distinct from other areas of science and which demand a different technological response. These challenges are: i) discovering interdependencies between disparate datasets representing different real-world phenomena and how one phenomenon can positively or negatively impact the other; ii) geospatial data integration and reasoning enabling environmental scientists to respond to the emerging trends or geographic events in a timely manner; iii) data heterogeneity that arises from using a wide variety of data formats, models, instruments and procedures; iv) data discovery and access problems that arise owing to the issues of geographically scattered environmental

data, temporally sparse data, restricted access to numerical models, institutional hindrance to data access for instance, due to compatibility issues or financial hurdles; v) data quality and provenance issues that arise due to many factors including faulty instruments, lack of metadata, naïve data collectors, bad environmental conditions and uneven practices. (Note that this thesis elected to focus on the first three challenges to maintain a clear scope in the work.)

### 6.3.2 Current Practices in Environmental Science

The thesis also contributes some insights into current practices in data management in environmental science, including an important exploration of technological opportunities and barriers. Some of the insights that have emerged from this work include:

- The practices and technologies are clearly insufficient and suffer from either methodological limitations (old technologies) or technical and financial issues (particularly related to environmental sensors and IoT technology).
- Because of the advanced measurement instruments that generate more data, there is now a trend towards more data-driven science to look for interesting and emergent patterns among different datasets and turning them into knowledge.
- Getting a unified view of the structure and more importantly semantics of complex and heterogeneous environmental data is very important.

Perhaps the most important result from this study though is the need for cross-disciplinary dialogue between environmental science and computer science so that technological opportunities can be delivered and barriers overcome.

### 6.3.3 Role of Semantic Web Technologies in Environmental Science

The most important set of contributions relate to an understanding of how semantic web technologies can support environmental science and in particular the unique data challenges in this area. Through the iterative development of an ontology for streaming environmental data, it has been shown that semantic web technologies have a significant role to play in overcoming these challenges. In particular, it has been shown how:

- Interdependencies between disparate datasets have been overcome by semantically enriching those low-level sensor measurements using the ontology and then reasoning over the resultant enriched datasets deriving new knowledge and interrelationships using associated inference rules, e.g. deducing a pollution event in the catchment.
- Geospatial data integration and reasoning issue have been resolved by again semantically enriching all sensor measurements using the ontology which has support for topological functions and predicates.
- Interoperable metric units conversion has been addressed by semantically assigning all sensor measurements their associated metric units using the ontology and then performing unambiguous translation between different metric units through inference rules.

The overall ontology is also a contribution in its own right providing a proof of concept of how a given ontology can address the needs for a given environmental project, in this case dealing with streaming data from an Environmental Internet of Things deployed in North Wales (the Conwy catchment). A set of principles underpin this design, namely re-use of existing ontologies where possible, the need for a modular approach, and the importance of having lightweight and relatively minimal ontologies which can develop over time.

### 6.3.4 Implications for Technological Infrastructure

The experimental work in this thesis has provided extra insights into the technological needs of environmental science and in particular the underlying infrastructure needed to support scientific discovery. One of the important insights is that existing infrastructure already exists in the form of software stacks supporting the Semantic Web and these can largely be adopted to support this particular area of science. In particular, this thesis has shown how existing technologies including ontologies, RDF, OWL, linked data and SPARQL can successfully be used in this domain. There are however some additional challenges that need to be met including issues around real-time data, with this being revisited in future work below.

### 6.3.5 Research Questions Revisited

The research questions from chapter one where as follows:

- What are the particular characteristics of data associated with environmental science, and what are the associated data challenges in terms of making sense of that data?
- What is the role of Semantic Web technologies in building a data model for the Environmental IoT Infrastructure to represent its data in all its complexity?
- What implications does this have for a technological infrastructure underpinning environmental science to exploit the potential of streaming data from IoT technology?

It should now be apparent that there is a strong mapping between the contributions and the initial research questions, namely that contributions 1 (section 6.3.1 entitled 'Characteristics of Environmental Data') and 2 (section 6.3.2 entitled 'Current Practices in Environmental Science') are in response to the first research question, with the following two sets of contributions (section 6.3.3 entitled 'Role of Semantic Web Technologies in Environmental Science' and section 6.3.4 entitled 'Implications for Technological Infrastructure') addressing the second and third research questions respectively.

## 6.4 Future Work

Some key areas of future research emanating from this research are outlined below.

### 6.4.1 Real-time Streaming Data

Most of the approaches presented in Semantic Web research for IoT/streaming data in supporting environmental science are based on using RDF data that is already stored in a database or a triplestore. However, IoT devices deployed in the natural environment will often be capturing data that need to be processed on-the-fly to respond to critical events in the environment. Hence, there is a need to develop Semantic Web techniques to reason over real-time steaming data. Reasoning over large spatio-temporal streaming data in an efficient and scalable manner is a huge challenge and hence is a key avenue for future work.

### 6.4.2 Bringing Together Ontology Development and Machine Learning

There is a potential symbiotic relationship between ontologies and machine learning. Ontologies can leverage from machine learning algorithms to add a probabilistic component to knowledge bases. Thus, ontologies can be informed through machine learning results. Similarly, a machine learning approach can incorporate existing ontologies to guide machine learning methods and learn new ways to extend the learning models. Hence novel techniques of data science could be developed through combining the advantages of these two prominent areas.

### 6.4.3 Semantic Web for Early Warning Systems

Natural disasters such as floods, hurricanes and tsunamis etc. can have major impact on human lives and economies. In order to reduce the effects of such disasters preventing human lives, an early warning system (EWS), based on IoT technologies is required to capture rich sets of ubiquitous real-time data. Such systems further require context and situation awareness to predict effectively such environmental hazards. Semantic Web technologies can play a potentially important role in understanding context awareness and reasoning over such environments. An early warning system, based on ontologies, semantic web services and semantic middleware, needs to be designed, which would be driven by semantically-enriched and dynamically constructed metadata. Such semantic computing models can potentially be able to predict environmental hazards in a timely manner.

### 6.4.4 Addressing the Uncertainty Challenge

Representing and managing uncertainty in earth and environmental science is a huge challenge. One of the main concerns of environmental scientists about uncertainty is the reasoning support in complex modelling scenarios, for instance, reasoning about propagating uncertainty in integrated modelling [177]. Uncertainty can arise from many sources: i) the underlying unreliable data sources methods, for instance, using citizen science, cheap and less reliable sensors, and lower satellite observations, for data collections; ii) choosing different models in experiments. In order to deal with uncertainty using Semantic Web technologies, a knowledge representation mechanism, i.e. ontology, is required to conceptualise and tackle the effects of

uncertain phenomena. However, current Semantic Web technologies do not have the ability to describe and reason over uncertainty in a principled way. Hence, there is a need to carry out further research to develop probabilistic ontologies providing a basis for reasoning to resolve the uncertainty challenge.

## 6.5 Final Remarks

The author argues that the approach presented in this thesis has the potential to address important data challenges. In particular, this thesis has examined the potential role of Semantic Web technologies for IoT/streaming data in underpinning environmental science and dealing with the associated challenges. The results presented in this work demonstrate the applicability and potential of such techniques, while also pointing several research avenues for further investigation. Finally, this thesis has examined the unique characteristics of environmental science around data in all its complexity. The author invites the Semantic Web research community to further explore the potential of these technologies and help environmental scientists to revolutionise their science.

# 7 References

[1] J. Han, J. Gao, “Research Challenges for Data Mining in Science and Engineering”, in Next Generation of Data Mining, Taylor and Francics Group LLC 2008.

[2] G. Bell, T. Hey, A. SzalayBeyond the Data Deluge. Science, 323 (5919) (2009), pp. 1297-1298

[3] Heidorn, P. B. 2008 Shedding light on the dark data in the long tail of science. Libr. Trends 57, 280-299. (doi:10.1353/lib.0.0036)

[4] Zoback ML. Grand challenges in Earth and Environmental Sciences – Science stewardship, and service for the twenty-first century. GSA Today 2001;11(12):41–7.

[5] A. Thorpe, Environmental eScience. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 367 (1890) (Mar. 2009), pp. 801-802

[6] S. Fang et al., "An integrated system for regional environmental monitoring and management based on Internet of Things", IEEE Trans. Ind. Informat., vol. 10, no. 2, pp. 1596-1605, May 2014.

[7] Hart, J. K., and K. Martinez (2006), Environmental sensor networks: A revolution in the Earth system science? Earth Sci. Rev., 78, 177–191.

[8] J. Fan, F. Han and H. Liu, "Challenges of Big Data analysis," National Science Review, in press, 2014.

[9] Tim Berners-Lee, Jim Hendler, and Ora Lassila, “The Semantic Web”, Scientific American, 284(5): May 2001, pp. 34-43.

[10] W. Wang, P. Bamaghi, "Semantic annotation and reasoning for sensor data", Proceedings of the 4th European conference on Smart sensing and context (EuroSSC2009), Springer-Verlag, 2009.

[11] http://www.w3.org/standards/semanticweb/data

[12] Bizer, C., T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1--22, 2009.

[13] Gruber, T, 1993. "A translation approach to portable ontology specifications". Knowledge Acquisition 5(2) 199-220.

[14] The Science Environment for Ecological Knowledge (SEEK) project, https://www.nceas.ucsb.edu/projects/6680

[15] L. Yu, Y. Liu, Using the linked data approach in a heterogeneous sensor web: challenges, experiments and lessons learned, In: Workshop on Sensor Web Enablement, 2011.

[16] J. Zhao, C. A. Goble, R. Stevens, and S. Bechhofer, "Semantically Linking and Browsing Provenance Logs for E-science," in ICSNW, 2004.

[17] C. Berkley, S. Bowers, M. Jones, J. Madin, and M. Schildhauer, "Improving data discovery for metadata repositories through semantic search". In Proc. of the Intl. Conf. on Complex, Intelligent and Software Intensive Systems (CISIS), 2009, pp. 1152–1159.

[18] Pouchard, L.C. et al, "A Linked Science Investigation: Enhancing Climate Change Data Discovery with Semantic Technologies". Earth Science Informatics 6, no. 3 (September 1, 2013): 175-85.

[19] Missier, P., Ludascher, B., Bowers, S., Dey, S., Sarkar, A., Shrestha, B., Altintas, I., Anand, M. K. and Goble, C. (2010). "Linking multiple workflow provenance traces for interoperable collaborative science". 5th Workshop on Workflows in Support of Large-Scale Science, New Orleans, USA.

[20] http://www.environmental-iot.com/

[21] N. Guarino, "Formal Ontology in Information Systems", Formal Ontology in Information Systems, pp. 3-15, 1998.

[22] Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved July 23, http://www.w3.org/DesignIssues/LinkedData.html

[23] P. Fox and J. Hendler. Semantic eScience: Encoding meaning in next-generation digitally enhanced science. The Fourth Paradigm: Data Intensive Scientific Discovery, Eds. Tony Hey, Stewart Tansley and Kristin Tolle, Microsoft External Research, pages 145–150, 2009.

[24] Fox, P., McGuinness, D.L., Raskin, R., Sinha, K., 2007. A volcano erupts: semantically mediated integration of heterogeneous volcanic and atmospheric data. In: Proceedings of the First Workshop on Cyberinfrastructure: Information Management in eScience, Lisbon, Portugal, November 9, 2007.

[25] McGuinness, D.L., Fikes, R., Rice, J. and Wilder, S. (2000). An Environment for Merging and Testing Large Ontologies. Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000). A. G. Cohn, F. Giunchiglia and B. Selman, editors. San Francisco, CA, Morgan Kaufmann Publishers.

[26] The Marine Metadata Interoperability (MMI) project, http://marinemetadata.org

[27] The Semantic Web Health Care and Life Sciences (HCLS) Interest Group https://www.w3.org/2001/sw/hcls/

[28] Good, B.M., Wilkinson, M.D. "The life sciences semantic web is full of creeps!", Brief Bioinform, 7 (2006), pp. 275-286

[29] Huajun Chen, Yimin Wang, Kei-Hoi Cheung (Eds.): Semantic e-Science**.** Annals of Information Systems 11, Springer 2010, ISBN 978-1-4419-5902-7

[30] Emmott S, Rison S. Towards 2020 Science Report. Microsoft Research, Cambridge, 2006.

[31] Hey, T. and Trefethen, A. E. 2005. Cyberinfrastructure for e-science. Science, 308(5723): 817–821.

[32] W.A. Wulf. The collaboratory opportunity. Science, 261 (1993), pp. 854-855

[33] Taylor, J. 2001. e-Science definition. http://www.e-science.clrc.ac.uk.

[34] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al., (2003). Revolutionizing science and engineering through cyberinfrastructure: Final report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.

[35] NSF Cyberinfrastructure Council (2006). NSF's cyberinfrastructure vision for 21st century discovery (Version 7.1). Retrieved from http://www.nsf.gov/od/oci/ci-v7.pdf

[36] Hall, W., De Roure, D., & Shadbolt, N. (2009). The evolution of the web and implications for eResearch. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367 (1890), 991.

[37] Lim H.B., Iqbal M., Yao Y., Wang W. (2010) A Smart e-Science Cyberinfrastructure for Cross-Disciplinary Scientific Collaborations. In: Chen H., Wang Y., Cheung KH. (eds) Semantic e-Science. Annals of Information Systems, vol 11. Springer, Boston, MA

[38] Eamon, W. “Science and the secrets of nature: Books of secrets in medieval and early modern culture”, 1994.

[39] Ziman, JM., “Public knowledge: An essay concerning the social dimension of science”, 1968.

[40] Horizon 2020 Consultation Report: Open Infrastructures for Open Science.

[41] Neylon C., Wu S. 2009 Open science: tools, approaches, and implications Pacific symposium on biocomputing 2009, 14th Symp, Hawaii, 5–9 January Altman R. B., Dunker A. K., Hunter L., Murray T., Klein 540–544 Singapore World Scientific (http://psb.stanford.edu/psb-online/proceedings/psb09/abstracts/2009_p540.html)

[42] G. Boulton, P. Campbell, B. Collins, P. Elias, W. Hall, G. Laurie, O. O'Neill, M. Rawlins, J. Thornton, P. Vallance and M. Walport, Science as an open enterprise, The Royal Society, 2012.

[43] Hey, T. (2010). The next scientific revolution. Harvard Business Review, 88(11), 56–63.

[44] Hey, T., Tansley, S., Tolle, K M, "The fourth paradigm: data-intensive scientific discovery", (2009).

[45] http://www.w3.org/Talks/WWW94Tim/

[46] W. Hall, “The Ever-Evolving Web: The Power of Networks,” Int’l J. Commun., vol. 5, 2011, p. 14.

[47] S. Decker, P. Mitra, S. Melnik, "Framework for the Semantic Web: An RDF Tutorial", *IEEE Internet Computing*, vol. 4, no. 6, pp. 68-73, Nov./Dec. 2000.

[48] O. Lassila, "Introduction to RDF Metadata", 1997. Available: https://www.w3.org/TR/NOTE-rdf-simple-intro

[49] Klyne, G. and Carroll, J. 2004. Resource description framework (RDF) concepts and abstract syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[50] E. Miller, An introduction to the resource description framework, Dlib Magazine, May 1998, ISSN 1082-9873, http://www.dlib.org/dlib/may98/miller/05miller.html

[51] Genesereth, M. R. and Nilsson, N. J. 1987. Logical Foundation of Artificial Intelligence. Morgan Kaufmann, Los Altos, California.

[52] Borst, W.N., "Construction of Engineering Ontologies". (2nd ed.), PhD Thesis, University of Twente, Enschede (1997)

[53] Studer R., Benjamins V.R., and Fensel, D. "Knowledge engineering, principles and methods". Data and Knowledge Engineering, 25(1-2):161–197, 1998.

[54] McCarthy, J. "Circumscription—a form of non-monotonic reasoning", Artif. Intell., 13 (1980), pp. 27-39

[55] Hayes, P. J. Naive Physics I: Ontology for liquids. Working Paper 63, lnstitut pour les Etudes Semantiques et Cognitives, Geneva, 1978.

[56] Alexander JH, Freiling MJ, Shulman SJ, Staley JL, Rehfuss S and Messick SL, 1986. “Knowedge level engineering: ontological analysis”. AAAI-86 21 963–968.

[57] Chandrasekaran, B. Josephson J. R. and Benjamins, R. (1999). What are ontologies, and why do we need them? IEEE Intelligent Systems, Vol.14, No.1, 20-26.

[58] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, 2001.

[59] SWEET (Semantic Web for Earth and Environmental Terminology), available from http://sweet.jpl.nasa.gov/ontology/

[60] Van Heijst, G., Schreiber, A. T., and Wielinga, B. J. 1997. Using Explicit Ontologies in KBS Development. International Journal of Human and Computer Studies, 46: 183-292.

[61] C. Roussey, F. Pinet, M. Kang, O. Corcho, An introduction to ontologies and ontology engineering, in: Ontologies in Urban Development Projects, Springer, London, 2011, pp. 9–38.

[62] Gruber, T. R. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies, 43(5/6): 907-928.

[63] Bard, J. B. & Rhee, S. Y. Ontologies in biology: design, applications and future challenges. Nature Rev. Genet. 5, 213–222 (2004).

[64] Twigger, S. et al. Rat Genome Database (RGD): mapping disease onto the genome. Nucleic Acids Res. 30, 125–128 (2002).

[65] Garcia-Hernandez, M. et al. TAIR: a resource for integrated Arabidopsisdata. Funct. Integr. Genomics 2, 239–253 (2002).

[66] Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E. & Brendel, V. MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res. 32, D393–D397 (2004).

[67] Gordon S. Blair et al, The role of ontologies in emergent middleware: supporting interoperability in complex distributed systems, Proceedings of the 12th ACM/IFIP/USENIX international conference on Middleware, December 12-16, 2011, Lisbon, Portugal [doi>10.1007/978-3-642-25821-3_21].

[68] D. L. McGuinness. 2002. Ontologies Come of Age. In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.

[69] Ben Mustapha, N. et al, "Combining Semantic Search and Ontology Learning for Incremental Web Ontology Engineering", Sixth International Workshop on Web Information Systems Modeling (WISM 2009) held in conjunction with CAISE'09, June 2009.

[70] Uschold, M., Gruninger, M., Ontologies and semantics for seamless connectivity, ACM SIGMOD Record, v.33 n.4, December 2004 [<doi>10.1145/1041410.1041420], Uschold, Gruninger.

[71] LUKASIEWICZ, T. 2011. Ontology-Based Semantic Search on the Web.

[72] Jasper, R. and Uschold, M. 1999. A Framework for Understanding and Classifying Ontology Applications. In Twelfth Workshop on Knowledge Acquisition Modelling and Management KAW'99 Applications.

[73] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 2011.

[74] http://microformats.org

[75] C. Bizer. The emerging web of linked data IEEE Intell. Syst., 24 (5) (2009), pp. 87-92.

[76] http://www.programmableweb.com

[77] Auer, S.; Lehmann, J.; Ngomo, A.-C. N.; and Zaveri, A. 2013. Introduction to linked data and its lifecycle on the web. In Reasoning Web. Semantic Technologies for Intelligent Data Access. Springer. 1–90.

[78] Berners-Lee T, Chen Y, Chilton L, Connolly D, Dhanaraj R, Hollenbach J, et al. Tabulator: exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06) workshop, Athens, Georgia, 6 November 2006.

[79] Marble (FU Berlin, Becker and Bizer) 2008.

[80] Disco Hypermedia Browser (http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/)

[81] Hastrup, T., Cyganiak, R., Bojars, U. (2008). Browsing Linked Data with Fenfire. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008) at 28th British National Conference on Databases (BNCOD2011).

[82] Kobilarov, G. and Dickinson, I. (2008). Humboldt: Exploring Linked Data, Proc. Linked Data on the Web (LDW'08), Beijing, China, April 2008.

[83] Gong Cheng , Weiyi Ge , Yuzhong Qu, Falcons: searching and browsing entities on the semantic web, Proceedings of the 17th international conference on World Wide Web, April 21-25, 2008, Beijing, China [doi>10.1145/1367497.1367676]

[84] Harth, A., Hogan, A., Delbru, R., Umbrich, J., O'Riain, S., and Decker, S. "SWSE: Answers before links!", In: Semantic Web Challenge, 2007.

[85] T. Finin et al., (2005) "Swoogle: Searching for knowledge on the Semantic Web", proceedings of the Twentieth National Conference on Artificial Intelligence.

[86] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies, 3(1):37–52, 2008.

[87] M. d'Aquin, M. Sabou, E. Motta, S. Angeletou, L. Gridinoc, V. Lopez, and F. Zablith. What can be done with the semantic web? an overview Watson-based applications. In SWAP, 2008.

[88] C. Becker, C. Bizer, "DBpedia Mobile: A Location-Enabled Linked Data Browser", Proc. World Wide Web 2008 Workshop: Linked Data on the Web (LDOW 08), 2008.

[89] Heath, T. and Motta, E., "Revyu: Linking reviews and ratings into the Web of Data", Web Semantics: Science, Services and Agents on the World Wide Web, 6(4):266–273, 2008.

[90] C. Clarke. A resource list management tool for undergraduate students based on linked open data principles. In Proceedings of the 6th European Semantic Web Conference, Heraklion, Greece, 2009.

[91] Sören Auer, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, "Introduction to linked data and its lifecycle on the web", Proceedings of the 7th international conference on Reasoning web: semantic technologies for the web of data, p.1-75, August 23-27, 2011, Galway, Ireland.

[92] Alberto Cerpa, Jeremy Elson, Michael Hamilton, Jerry Zhao, Deborah Estrin, Lewis Girod, Habitat monitoring: application driver for wireless communications technology, Workshop on Data communication in Latin America and the Caribbean, p.20-41, April 2001, San Jose, Costa Rica [<doi>10.1145/371626.371720]

[93] J. M. Kahn, R. H. Katz, and K. S. J. Pister, "Next century challenges: Mobile networking for smart dust," In Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, ser. MobiCom '99. New York, NY, USA: ACM, 1999, pp. 271-278.

[94] Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, John Anderson, Wireless sensor networks for habitat monitoring, Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications, September 28-28, 2002, Atlanta, Georgia, USA.

[95] Robert Szewczyk, Alan Mainwaring, Joseph Polastre, John Anderson, David Culler, An analysis of a large scale habitat monitoring application, Proceedings of the 2nd international conference on Embedded networked sensor systems, November 03-05, 2004, Baltimore, MD, USA.

[96] Philo Juang, Hidekazu Oki, Yong Wang, Margaret Martonosi, Li Shiuan Peh, Daniel Rubenstein, "Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet", Proceedings of the 10th international conference on Architectural support for programming languages and operating systems, October 05-09, 2002, San Jose, California.

[97] Robert Szewczyk, Eric Osterweil, Joseph Polastre, Michael Hamilton, Alan Mainwaring, Deborah Estrin, Habitat monitoring with sensor networks, Communications of the ACM, v.47 n.6, June 2004.

[98] Martinez, K., Ong, R. and Hart, J. (2004) "Glacsweb: a sensor network for hostile environments" In Proc. of the First IEEE SECON Conference 2004, Santa Clara, USA.

[99] W.L. Lee, A. Datta, R. Cardell-Oliver, Network Management in Wireless Sensor Networks, [online] Available: http://www.csse.uwa.edu.aurwinnielNetwork_Management_in_WSNs.pdf.

[100] Gilman Tolle, Joseph Polastre , Robert Szewczyk , David Culler , Neil Turner , Kevin Tu , Stephen Burgess , Todd Dawson , Phil Buonadonna , David Gay , Wei Hong, A macroscope in the redwoods, Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems, November 02-04, 2005, San Diego, California, USA.

[101] K. Langendoen, A. Baggio, and O. Visser. Murphy loves potatoes: Experiences from a pilot sensor network deployment in precision agriculture. In 14th Int. Workshop on Parallel and Distributed Real-Time Systems (WPDRTS), pages 1--8, apr 2006.

[102] G. Werner-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees, M. Welsh, "Deploying a Wireless Sensor Network on an Active Volcano", IEEE Internet Comput., vol. 10, no. 2, pp. 18-25, March-April 2006.

[103] Andreas Terzis , Razvan Musaloiu-E. , Joshua Cogan , Katalin Szlavecz , Alexander Szalay , Jim Gray , Stuart Ozer , Chieh-Jan Mike Liang , Jayant Gupchup , Randal Burns, Wireless sensor networks for soil science, International Journal of Sensor Networks, v.7 n.1/2, p.53-70, February 2010.

[104] G. Bishop-Hurley, D. Swain, D. Anderson, P. Sikka, C. Crossman, P. Corke, "Virtual fencing applications: Implementing and testing an automated cattle control system", Comput. Electron. Agriculture, vol. 56, no. 1, pp. 14-22, Mar. 2007.

[105] P. Grace, D. Hughes, B. Porter, G. Blair, G. Coulson, and F. Taiani, "Experiences with open overlays: a middleware approach to network heterogeneity," in Proc. of the 3rd ACM SIGOPS/EuroSys European Conference on Computer Systems 2008, April 2008, pp. 123-136.

[106] L. Mo, Y. He, Y. Liu, J. Zhao, S.-J. Tang, X.-Y. Li, and G. Dai, \Canopy closure estimates with greenorbs: Sustainable sensing in the forest," in Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, ser. SenSys '09. New York, NY, USA: ACM, 2009, pp. 99{112.

[Online] Available: http://doi.acm.org/10.1145/1644038.1644049

[107] F. Ingelrest, G. Barrenetxea, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, "Sensorscope: Application-specic sensor network for environmental monitoring," ACM Trans. Sen. Netw., vol. 6, no. 2, pp. 17:1-17:32, Mar. 2010. [Online] Available: http://doi.acm.org/10.1145/1689239.1689247

[108] Corke, T.Wark, R. Jurdak, W. Hu, P. Valencia, and D. Moore, "Environmental wireless sensor networks," Proceedings of the IEEE, vol. 98, no. 11, pp. 1903-1917, Nov 2010.

[109] https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf

[110] S. Avancha, C. Patel, and A. Joshi. Ontology-driven adaptive sensor networks. In 1st Annual International Conference on Mobile and Ubiquitous Systems, Networking and Services, 2004.

[111] D. Russomanno, C. Kothari, and O. Thomas. Building a sensor ontology: a practical approach leveraging ISO and OGC models. In 2005 International Conference on Artificial Intelligence (vol 2), 2005.

[112] D. Russomanno, C. Kothari, and O. Thomas. Sensor ontologies: from shallow to deep models. In 37th Southeastern Symposium on System Theory, 2005.

[113] J. Kim, H. Kwon, D. Kim, H. Kwak, and S. Lee. “Building a service-oriented ontology for wireless sensor networks”, in Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science, 2008, pp. 649-654.

[114] K. J. Witt, J. Stanley, D. Smithbauer, D. Mandl, V. Ly, A. Underbrink, and M. Metheny. "Enabling Sensor Webs by utilizing SWAMO for autonomous operations". In 8th NASA Earth Science Technology Conference, 2008.

[115] A. Herzog, D. Jacobi, and A. Buchmann. A3ME - an Agent-Based middleware approach for mixed mode environments. In 2nd International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008.

[116] A. Herzog, D. Jacobi, and A. Buchmann. Predefined classification for mixed mode environments. Technical report, TU Darmstadt, 2009.

[117] M. Compton, H. Neuhaus, and K.-N. Tran. Reasoning about sensors and compositions. In 2nd International Semantic Sensor Networks Workshop, 2009.

[118] H. Neuhaus and M. Compton. The semantic sensor network ontology: a generic language to describe sensor assets. In AGILE Workshop: Challenges in Geospatial Data Harmonisation, 2009.

[119] A. Sheth, C. Henson, and S. Sahoo. Semantic sensor web. IEEE Internet Computing, 12(4), 2008.

[120] W. Kuhn, A functional ontology of observation and measurement, in: Proceedings of the 3rd International Conference on GeoSpatial Semantics, SpringerVerlag, 2009, pp. 26–43.

[121] M. Compton, P.M. Barnaghi, L. Bermudez, R. Garcia-Castro, Ó Corcho, S. Cox, J. Graybeal, M. Hauswirth, C.A. Henson, A. Herzog, V.A. Huang, K. Janowicz, W.D. Kelsey, D.L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K.R. Page, A. Passant, A.P. Sheth, K. Taylor, The SSN ontology of the W3C semantic sensor network incubator group, J. Web Semant. 17 (2012) 25–32.

[122] https://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/

[123] K. Janowicz, M. Compton, The stimulus-sensor-observation ontology design pattern and its integration into the semantic sensor network ontology, in: 3rd International Workshop on Semantic Sensor Networks, vol. 668, CEUR-WS, 2010.

[124] A. J. Gray, R. García-Castro, K. Kyzirakos, M. Karpathiotakis, J.-P. Calbimonte, K. Page, et al., "A semantically enabled service architecture for mashups over streaming and stored data", in The Semantic Web: Research and Applications, ed: Springer, (2011), pp. 300-314.

[125] Wang, W.; De, S.; Cassar, G.; Moessner, K. Knowledge representation in the internet of things: Semantic modelling and its applications. Automatika-J. Control Meas. Electron. Comput. Commun. 2013, 54, 388–400.

[126] P. Barnaghi, F. Ganz, C. Henson, A. Sheth, "Computing perception from sensor data", Proc. IEEE Sensors, pp. 1-4, 2012.

[127] Roda, F., Musulin, E., (2014), "An ontology-based framework to support intelligent data analysis of sensor measurements". Expert Systems with Applications 41, 7914–7926.

[128] K. Taylor, C. Griffith, L. Lefort, R. Gaire, M. Compton, T. Wark, D. Lamb, G. Falzon, M. Trotter, "Farming the web of things", Intelligent Systems IEEE, vol. 28, no. 6, pp. 12-19, 2013.

[129] P. Barnaghi, S. Meissner, M. Presser, and K. Moessner. Sense and sens'ability: semantic data modelling for sensor networks. In ICT Mobile Summit, 2009.

[130] A. Bröring, P. Maué, K. Janowicz, D. Nüst, C. Malewski, "Semantically-enabled sensor plug & play for the sensor web", Sensors, 11 (2011), pp. 7568-7605

[131] V. Huang and M. K. Javed, "Semantic sensor information description and processing," in 2nd International Conference on Sensor Technologies and Applications, SENSORCOMM 2008, 2008, pp. 456-461.

[132] Moraru, A. and Mladenic, D. (2012). A framework for semantic enrichment of sensor data. In Information Technology Interfaces (ITI), Proceedings of the ITI 2012 34th International Conference on, pages 155–160.

[133] M. Compton, C. Henson, H. Neuhaus, L. Lefort, A. Sheth, "A survey of the semantic specification of sensors", 2nd International Workshop on Semantic Sensor Networks at 8th International Semantic Web Conference", Oct. 2009.

[134] Villa F, Athanasiadis IN, Rizzoli AE (2009) Modelling with knowledge: a review of emerging semantic approaches to environmental modelling. Environmental Modelling & Software 24: 577–587.

[135] Porter, J. H.2010b. A controlled vocabulary for LTER datasets. 2010. http://databits.lternet.edu/spring-2010/controlled-vocabulary-lter-datasets

[136] Raskin, R., 2005. Semantic web for earth and environmental terminology (SWEET) [online: https://sweet.jpl.nasa.gov/].

[137] Tripathi A, Babaie HA (2008) Developing a modular hydrogeology ontology by extending the sweet upper-level ontologies. Comput Geosci 34(9):1022–1033.

[138] Williams R., Martinez N., Golbeck J. 2006 Ontologies for ecoinformatics, Journal of web semantics, Web Semantics: Science, Services and Agents on the World Wide Web 4 (2006) 237–242.

[139] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. Ecological Informatics, 2(3):279–296, October 2007.

[140] Van Der Werf, B., Adamescu, M., Ayromlou, M., Bertrand, N., Borovec, J. and Frenzel, M. 2009. "A long-term biodiversity, ecosystem and awareness research network".

[141] Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, et al. (2013) The environment ontology: contextualising biological and biomedical entities. J Biomed Semantics 4: 43.

[142] Buttigieg P.L.Pafilis E.Lewis S.E. et al., (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J. Biomed. Semant., 7, 57.

[143] Viral Parekh, J.-P.J.G., and Tim Finin. "Ontology based Semantic Metadata for Geoscience Data", International Conference of Information and Knowledge Engineering. June 21, 2004.

[144] Michener WK and Jones MB. 2011. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecol and Evol 27: 83–93.

[145] Fegraus EH, Andelman S, Jones MB, Schildhauer M (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation, Bulletin of the Ecological Society of America 86(3): 158–168.

[146] Z. Malik, A. Rezgui, A. K. Sinha, K. Lin, and A. Bouguettaya 2007, DIA: A Web Services-based Infrastructure for Semantic Integration in Geoinformatics, Proceedings of the IEEE ICWS 2007, Application Services and Industry Track, submitted.

[147] Leinfelder, B., S. Bowers, M. O'Brien, M. B. Jones, M. Schildhauer. Using semantic metadata for discovery and integration of heterogeneous ecological data M.B. Jones, C. Gries (Eds.), Proceedings of the Environmental Information Management Conference (EIM 2011), University of California, pp. 92-97

[148] Bleisch, S., Duckham, M., Galton, A., Laube, P., Lyon, J.: Mining candidate causal relationships in movement patterns. Int. J. Geogr. Inf. Sci. 28(2), 363–382 (2013)

[149] O. J. Reichman, M. B. Jones, M. P. Schildhauer, "Challenges and opportunities of open data in ecology", Science, vol. 331, pp. 703-705, 2011.

[150] http://knb.ecoinformatics.org/software/eml/

[151] Ramachandran, R., S. Graves, H. Conover and K. Moe, "Earth Science Markup Language", Submitted to Computers & Geosciences Journal, Accepted with revisions, 2003.

[152] Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0. . (last accessed 23.01.08.).

[153] P. Fox, D. McGuinness, R. Raskin, and K. Sinha, A volcano erupts: Semantically mediated integration of heterogeneous volcanic and atmospheric data,

in: Proc. of the ACM First Workshop on CyberInfrastructure: Information Management in eScience, ACM, New York, NY, USA, 2007, pp. 1–6. doi:10.1145/1317353.1317355.

[154] T. Tarasova, M. Argenti, M. Marx. Semantically-Enabled Environmental Data Discovery and Integration: demonstration using the Iceland Volcano Use Case. P. Klinov and D. Mouromtsev (Eds.): KESW 2013, CCIS 394, pp. 289-297, 2013.

[155] Moura, A.M.D.C., Porto, F., Poltosi, M., Palazzi, D.C., Magalhaes, P., and Vidal, V., (2012). Integrating Ecological Data Using Linked Data Principles. In Joint V Seminar on Ontology Research in Brazil, pages 156–167.

[156] Mai, G.-S., Y.-H. Wang, Y.-J. Hsia, S.-S. Lu, and C.-C. Lin. 2011. Linked Open Data of Ecology (LODE): a new approach for ecological data sharing. Taiwan Journal of Forest Science: 26:417–424.

[157] A. Shaon, A. Woolf, S. Crompton, R. Boczek, W. Rogets, M. Jackson. An open source linked data framework for publishing environmental data under the UK location strategy, in: Terra Cognita Workshop in International Semantic Web Conference, ISWC2011.

[158] McLaren, R. and Waters, R. 2011. Governing Location Information in the UK. Cartographic Journal, The. 48, 3 (2011), 7.

[159] C. Rueda, N. Galbraith, R. A. Morris, L. E. Bermudez, R. A. Arko and J. Graybeal, “The MMI Device Ontology: Enabling Sensor Integration”, in AGU Fall Meeting Abstracts, **(2010)**, pp. 08.

[160] M. Calder, R. Morris, and F. Peri. Machine reasoning about anomalous sensor data. In International Conference on Ecological Informatics, 2008.

[161] G. Atemezing, O. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, et al., "Transforming meteorological data into linked data", Semantic Web, vol. 4, (2013), pp. 285-290.

[162] C. Henson, H. Neuhaus, A. Sheth, K. Thirunarayan, and R. Buyya. An ontological representation of time series observations on the semantic sensor web. In 1st International Workshop on the Semantic Sensor Web, 2009.

[163] Devaraju, A., Kauppinen, T.: Sensors Tell More than They Sense: Modeling and Reasoning about Sensor Observations for Understanding Weather Events. International Journal of Sensors, Wireless Communications and Control, Special Issue on Semantic Sensor Networks 2(1) (2012) ISSN: 2210-3279

[164] X. Su et al., "Connecting IoT sensors to knowledge-based systems by transforming SenML to RDF", Proc. 5th Int. Conf. Ambient Syst. Netw. Technol., pp. 215-222, Jun. 2014.

[165] K. Thirunarayan, C. Henson, and A. Sheth, "Situation Awareness via Abductive Reasoning from Semantic Sensor Data: A Preliminary Report," in The International Symposium on Collaborative Technologies and Systems, Maryland, USA, 2009, pp. 111-118.

[166] J. Yu and K. Taylor., "Event dashboard: Capturing user-defined semantics events for event detection over real-time sensor data". In Proc. Intl. Workshop on Semantic Sensor Networks, CEUR-WS Vol. 1063, pp. 35–50, October 2013.

[167] C. Roussey, S. Bernard, G. André, O. Corcho, G. De Sousa, D. Boffety, and J.-P. Chanet. Weather station data publication at irstea: an implementation report. In Joint Proceedings of the 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web and 7th International Workshop on Semantic Sensor Networks, volume 1401. CEUR, Oct. 2014.

[168] L. Lefort, J. Bobruk, A. Haller, K. Taylor, and A. Woolf. A Linked Sensor Data Cube for a 100-year homogenised daily temperature dataset. In C. A. Henson, K. Taylor, and Ó. Corcho, editors, Proceedings of the 5th International Workshop on Semantic Sensor Networks, SSN12, Boston, Massachusetts, USA, November 12,

2012, volume 904 of CEUR Workshop Proceedings, pages 1–16. CEUR-WS.org, 2012. http:// ceur-ws.org/Vol-904/paper10.pdf.

[169] Patricia Yancey Martin & Barry A. Turner, "Grounded Theory and Organizational Research," The Journal of Applied Behavioural Science, vol. 22, no. 2 (1986), 141.

[170] Fusch, P., & Ness, L. (2015). Are we there yet? Data saturation in qualitative research. The Qualitative Report, 20, 1208–1416.

[171] Strauss A. & Corbin J. Basics of Qualitative Research: Grounded Theory Procedures and Techniques, 1990. Sage, Newbury Park, CA.

[172] Strauss A. & Corbin J. Grounded theory methodology: An overview. In Handbook of Qualitative Research (Denzin N.K. & Lincoln Y.S. eds), Sage, Thousand Oaks, CA, 1994, pp. 273-285.

[173] Glaser B.G. & Strauss A.L. The Discovery of Grounded Theory. Aldine, Chicago, 1967.

[174] Glaser, B. G. 1978. Theoretical Sensitivity: Advances in the Methodology of Grounded Theory, Mill Valley, CA., Sociology Press.

[175] Charmaz K. The grounded theory method: An explication and interpretation. In Contemporary Field Research: A Collection of Readings (Emerson R.M. ed.), Waveland Press, Prospect Heights, IL, (1983) pp. 109-126.

[176] Chesler M.A. Professionals' Views of the Dangers of Self-help Groups: Explicating a Grounded Theoretical Approach (Center for Research on Social Organization, Working Paper Series). Department of Sociology, University of Michigan, Ann Arbour, (1987).

[177] Keith Beven and Rob Lamb. 2017. The uncertainty cascade in model fusion. In: Integrated environmental modelling to solve real world problems. Geological Society of London, London, pp. 255-266. http://dx.doi.org/10.1144/SP408.3

[178] Mackechnie, C., Maskell, L., Norton, L., and Roy, D. 2011. The role of 'Big Society' in monitoring the state of the natural environment. Journal of Environmental Monitoring 13, 10, 2687--2691.

[179] The "Bromley Principles" Regarding Full and Open Access to "Global Change" Data. By Allan Bromley, published in Policy Statements on Data Management for Global Change Research from the U.S. Office of Science and Technology Policy, July 2, 1991.

[180] http://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

[181] Christine L. Borgman, The conundrum of sharing research data, Journal of the American Society for Information Science and Technology, v.63 n.6, p.1059-1078, June 2012 [<doi>10.1002/asi.22634]

[182] http://www.computerweekly.com/opinion/The-problem-with-Open-Data

[183] Peter Buneman, Sanjeev Khanna, Wang Chiew Tan, "Why and Where: A Characterization of Data Provenance", Proceedings of the 8th International Conference on Database Theory, p.316-330, January 04-06, 2001.

[184] http://www.w3.org/TR/owl2-overview

[185] K. Sengupta and P. Hitzler, “Web ontology language (OWL),” in Encyclopedia of Social Network Analysis and Mining, 2014, pp. 2374– 2378. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-6170-8_ 113

[186] https://www.w3.org/TR/owl-features

[187] Mascardi V, Cord V, Rosso P (2006) A comparison of upper ontologies. Technical Report DISI-TR-06-21, Dipartimento di Informatica e Scienze dell'Informazione (DISI), Universita degli Studi di Genova, Italy.

[188] http://ontologydesignpatterns.org/ont/dul/DUL.owl

[189] http://www.loa.istc.cnr.it/old/DOLCE.html

[190] http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite

[191] A. Sheth and M. Perry, “Traveling the Semantic Web through Space, Time, and Theme,” IEEE Internet Computing, vol. 12, no. 2, 2008, pp. 81–86.

[192] http://www.w3.org/2003/01/geo/

[193] http://www.w3.org/2011/02/GeoSPARQL.pdf

[194]http://ontolog.cim3.net/file/work/EarthScienceOntolog/2012-12-12_EarthScienceOntolog_session-5/GeoSPARQL_Getting_Started--DaveKolas_20121212.pdf

[195] http://www.w3.org/TR/owl-time/

[196] P. Doran, V. Tamma, L. Iannone, Ontology module extraction for ontology reuse: an ontology engineering perspective, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management CIKM, 2007, pp. 61–70.

[197] http://graphdb.ontotext.com/

[198] I. Horrocks, P.F. Patel-Schneider, S. Bechhofer, D. Tsarkov, OWL rules: A proposal and prototype implementation, J. Web Sem. 3 (1) (2005) 23–40.

[199] http://www.w3.org/Submission/SWRL/

[200]http://protege.stanford.edu/conference/2009/slides/SWRL2009ProtegeConference.pdf

[201] https://www.w3.org/Submission/SWRL/#3.1

[202] http://jena.apache.org/documentation/inference/#rules

[203] R. Battle and D. Kolas. GeoSPARQL: Enabling a Geospatial Semantic Web. Semantic Web Journal, 3(4):355-370, 2012.

[204] Russ T, Valente A, MacGregor R, Swartout W (1999) Practical experiences in trading off ontology usability and reusability. In: Proceedings of the knowledge acquisition workshop, KAW99, Banff, Canada.