IntechOpen

Latest Advances and New Visions of Ontology in Information Science

Edited by Morteza SaberiKamarposhti and Mahdi Sahlabadi

Latest Advances and New Visions of Ontology in Information Science

Edited by Morteza SaberiKamarposhti and Mahdi Sahlabadi

Published in London, United Kingdom

Latest Advances and New Visions of Ontology in Information Science
http://dx.doi.org/10.5772/intechopen.102276
Edited by Morteza SaberiKamarposhti and Mahdi Sahlabadi

Contributors
G.N. Chiranjeevi, Subhash Kulkarni, Nicholas Nicholson, Iztok Štotl, Armita Davarpanah, Hassan A. Babaie, Guanyu Huang, Dionysia Varvarigou, David Espes, Giacomo Bersano, Yaseein Soubhi Soubhi Hussein, Maen Alrashdan, Ahmed Saeed Alabed, Saleh Alomar, Amjed Zraiqat

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2023 by IntechOpen
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, 5 Princes Gate Court, London, SW7 2QJ, United Kingdom

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Latest Advances and New Visions of Ontology in Information Science
Edited by Morteza SaberiKamarposhti and Mahdi Sahlabadi
p. cm.
Print ISBN 978-1-80356-917-8
Online ISBN 978-1-80356-918-5
eBook (PDF) ISBN 978-1-80356-919-2

Meet the editors

Morteza Saberi Kamarposhti received a Ph.D. in Computer Science from the University of Technology Malaysia. He is an assistant professor at the National University of Malaysia (UKM). He has several years of experience in teaching, research, administration, programming, and student affairs. Dr. Saberikamarposhti also has several research publications in well-known international journals and conferences to his credit.

Mahdi Sahlabadi holds a Ph.D. in Industrial Computing from the National University of Malaysia (UKM) and has worked as a researcher at several prestigious institutions, including the Japan Advanced Institute of Science and Technology (JAIST), Singapore Management University (SMU), and Sharif University of Technology (SUT). Currently, he is a post-doc researcher at UKM. Dr. Sahlabadi has made significant contributions to academia and industry and is known for his active involvement in these fields.

Contents

Preface

The word “anthology” originates from Latin, with its roots in two components: “anto,” meaning existence, and “logia,” meaning knowledge and study. Therefore, it can be associated with ontology or the study of existence. The first part of this combination refers to the realm of existence, while the second part denotes the knowledge that ancient Greeks sought to attain. Therefore, ontology emerged as a means for the Greeks to understand the world, hence its name. In the context of artificial intelligence, the term “ontology” describes the explicit meaning of the semantic web. It involves the comprehensive classification of objects and their connections in the universe. These concepts and ideas have their origins in deeper philosophical notions. Thomas Gruber, an influential figure in ontology, provided a notable definition for ontologies. According to him, an ontology is a formal representation of a set of terms and their relationships, expressed in a specialized language and stored in a computer-readable file. It is important to note that while “anthology” and “ontology” share a common root, they have distinct meanings.

“Anthology” refers to a collection of literary works, while “ontology” pertains to the formal representation of knowledge about a domain, particularly in the context of computer science and artificial intelligence.

Ontologies can be employed in several domains, such as global semantic networks, search engines, electronic commerce, natural language processing, knowledge engineering, information extraction and retrieval, multi-agent systems, qualitative modeling of physical systems, database design, information systems, and geographic and digital libraries.

The distinction between ontology in philosophy and ontology in computer science is fundamental. In philosophy, ontology emerges from the inherent order between concepts. However, in computer science, ontology is derived based on the order we assign to concepts. Furthermore, philosophical ontology aims for a comprehensive and universal perspective encompassing all concepts. Conversely, computer science ontology has a narrower focus, excluding elements outside the scope of discussion. It is worth noting that ontologies are not new to the Web. Each “cloud data design” can be seen as an ontology, as it defines a set of conceptual or physical attributes applicable to a specific user group. Typically, ontology is defined as the vocabulary of concepts in a particular field of knowledge organized in a hierarchical structure.

When discussing a topic, it is crucial to understand the subject clearly. Different fields use words with specific meanings, and sometimes even within a clear topic, people may interpret a word differently. The solution is to establish common words that everyone understands. This is where ontology comes into play on the Web.

An ontology lists all the entities in a domain, including their characteristics and connections. The information is organized in a specific format and linked to the Internet documentation. This helps establish standardized meanings. Ontologies are used by people, databases, and applications that need to share information about a particular domain.

The process of generating ontologies can be carried out manually using ontology engineering tools or through (semi-)automated methods of knowledge acquisition and construction. However, manual ontology building can be expensive, time-consuming, error-prone, and often reflects the personal opinions of the designer. Moreover, these manually built ontologies are typically inflexible and need more adaptability to specific changes for which they were developed. Automating the ontology construction process reduces costs and results in ontologies that are better suited to their intended applications. While ontology engineering tools serve as interfaces for application development, they still require human creators to utilize them effectively. However, by moving towards automating knowledge acquisition from various sources such as texts, databases, and existing ontologies, the challenges of ontology engineering can be addressed, leading to reduced costs and easier sharing of ontologies.

This book highlights the latest advancements and novel viewpoints in the field of ontology within information science.

Morteza SaberiKamarposhti
Assistant Professor,
Cyber Security Lab (CYBER),
Department of Computer Engineering,
Universiti Kebansaan Malaysia (UKM),
Bangi, Malaysia

Mahdi Sahlabadi
Post-doc Researcher,
Universiti Kebansaan Malaysia (UKM),
Bangi, Malaysia

Section 1

Infrastructures

Chapter 1

Data Centre Infrastructure: Design and Performance

Yaseein Soubhi Hussein, Maen Alrashd, Ahmed Saeed Alabed and Saleh Alomar

Abstract

The tremendous growth of e-commerce requires an increase in the data centre capacity and reliability for appropriate quality of services. Optimisation of data centre design is considered to be within a green technology that shows great promise to decrease CO_2 emission. However, a huge data centre requires huge power consumption due to higher capacity of racks that lead to more powerful cooling systems, power supply, protection and security. These make the data centre costly and not feasible for services. In this chapter, we will provide a tire 4 data centre design to be located in the optimal location of Malaysia, in Cyberjaya. The main purpose of this design is to provide e-commerce services, especially food delivery, with high quality of services and feasibility. All data centre components have been well designed to provide various services which include top-level security, colocation system, reliable data management and IT infrastructure management. Moreover, recommendation and justification have been provided to ensure that the proposed design outperforms compared to other data centres in terms of reliability, power effeminacy and storage capacity. In conclusion, analysing, synthesising and evaluating each component of the proposed data centre will be summarised.

Keywords: data centre, storage infrastructure, data centre infrastructure management (DCIM), security, scalability

1. Introduction

Meza is one of the home-grown data centre companies, and it provides various services which include top-level security and reliable data management and IT infrastructure management. Meza is expected to be built in several data centres across Malaysia. A Malaysian food delivery application company has more than 5 million users of the services, and the number of users is increasing day by day. The current infrastructure is insufficient to handle the vast amount of data processing, and it might cause the users to face poor user experience due to longer response time from the server and slow process. Therefore, the company has appointed Meza to construct a data centre to cater to the continued growth of the company. The data centre will be required to process online food ordering and online payment, customer relationship management to manage the communication in one inbox and engage with their client.

This chapter will be proposing a data centre design with the essential components for this food delivery application company and analysing, synthesising and evaluating each component of the proposed data centre. Other data centre components, such as power usage effectiveness and efficiency, cooling system and protection, have been discussed in the chapter Data Centre Infrastructure Power Efficiency and Protection (**Figure 1**).

2. Analysis

2.1 Customer requirements

The first basic requirement that comes to mind when building a data centre is the customer requirements. The customers are an essential entity. This data centre proposal has been designed to accept enormous amount of traffic loads, which means that many customers can order at the same time without the need to face frustration which happens when a system crashes due to overload. Furthermore, a seamless online chat function has been proposed which functions from the time of the order till the order has been delivered. This feature will enable customers to be more confident to use this delivery system as they can raise any issues regardless of the payment, order and so on. This is possible due to the new proposed network infrastructure.

Moreover, due to its newly improved network infrastructure, orders can be grouped more efficiently and can be delivered quickly. Lastly, the data centre is designed to be transparent which will allow both the customers and the delivery guy to key in ratings, which will compel the individuals to be reputable in order to achieve perks and hence make the system more trustworthy.

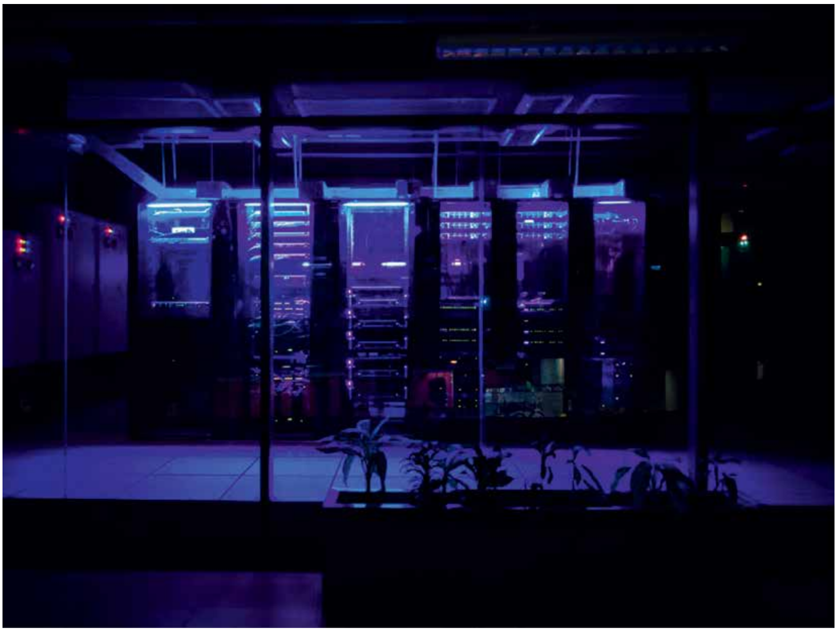


Figure 1.
APU data centre.

2.2 Data centre requirements

Moving on, there are a few requirements for a data centre to be robust such as:

- Availability/tier selection: to achieve high availability, Meza has decided to critically analyse between different types of data centre tiers. There are basically 4 types of tiers available. After comparing between the 4 types of tiers, the company has decided to go with tier 4 since the company is used for food delivery application and has a huge number of customers. As stated by [1], a tier 4 data centre has an uptime of 99.995% per year and has a '2 N + 1' completely redundant infrastructure which serves the sole purpose of the food delivery application. It has an annual downtime of only 26.3 minutes per year when compared to 1.6 hours per year for tier 3.
  Furthermore, tier 4 has been chosen for Meza because of the failure-tolerant design. As expressed by [2], failure-tolerant design is an essential part of the many benefits offered by a tier 4 data centre. This allows unplanned failures to be maintained that would otherwise cause critical loads in the site's infrastructure. Additionally, if any distribution or capacity component fails, the computer equipment of a tier 4 data centre shall not be affected. In order to avoid further disasters, the system will respond automatically. Moreover, there are also several distribution paths in a tier 4 data centre that can handle the computer equipment of the site at the same time. The IT equipment are all powered doubly and offer additional backup. Lastly, it is also supported by [3] that for mission-critical applications and systems, fault tolerance is especially crucial. This tier has the level of protection that is the most important, and tier 4 also provides support for electricity outage protection for 96 hours (**Figures 2** and **3**).

- Scalability: The food-ordering application company has more than 5 million active users and is increasing with time. The planned data centre will also be able to offer

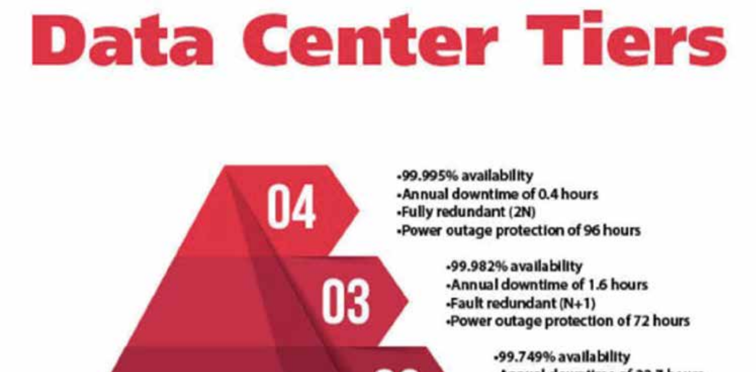
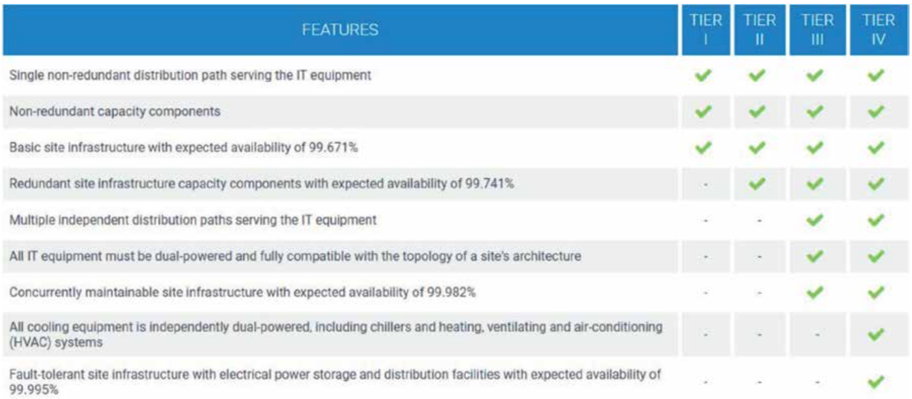

Data Center Tiers
04
•99.995% availability
•Annual downtime of 0.4 hours
•Fully redundant (2N)
•Power outage protection of 96 hours
03
•99.982% availability
•Annual downtime of 1.6 hours
•Fault redundant (N+1)
•Power outage protection of 72 hours
02
•99.749% availability
•Annual downtime of 22.7 hours
•Partial redundancy
•One path of power
01
•99.671% availability
•Annual downtime of 28.8 hours
•No redundancy
•One path of power and cooling

Figure 2.
Some features of different data centre tiers [4].



FEATURES	TIER I	TIER II	TIER III	TIER IV
Single non-redundant distribution path serving the IT equipment	✓	✓	✓	✓
Non-redundant capacity components	✓	✓	✓	✓
Basic site infrastructure with expected availability of 99.671%	✓	✓	✓	✓
Redundant site infrastructure capacity components with expected availability of 99.741%	-	✓	✓	✓
Multiple independent distribution paths serving the IT equipment	-	-	✓	✓
All IT equipment must be dual-powered and fully compatible with the topology of a site's architecture	-	-	✓	✓
Concurrently maintainable site infrastructure with expected availability of 99.982%	-	-	✓	✓
All cooling equipment is independently dual-powered, including chillers and heating, ventilating and air-conditioning (HVAC) systems	-	-	-	✓
Fault-tolerant site infrastructure with electrical power storage and distribution facilities with expected availability of 99.995%	-	-	-	✓

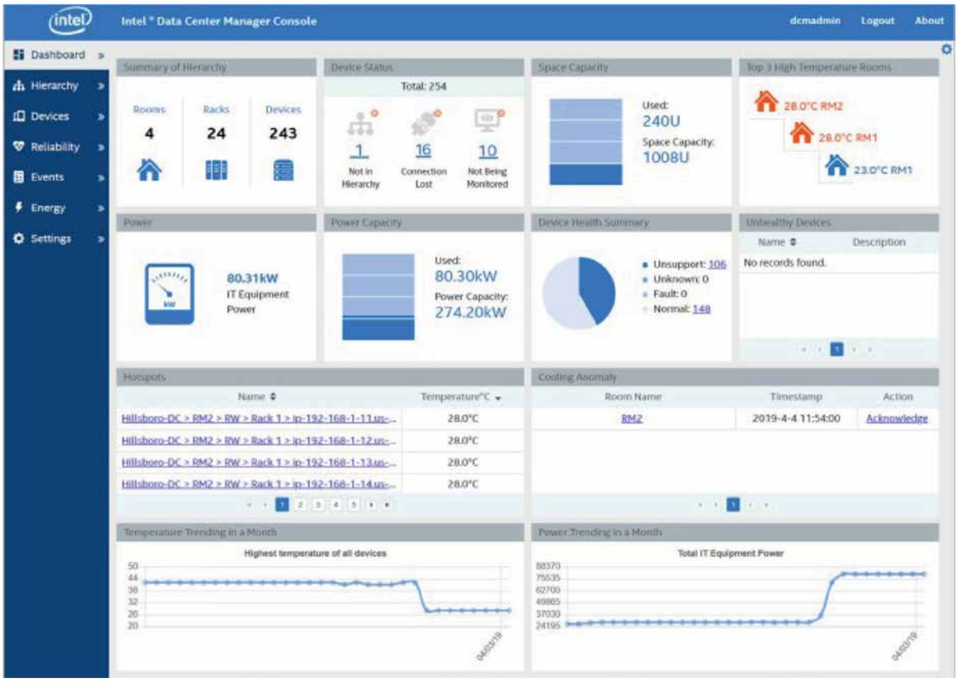
Figure 3.
Some features of different data centre tiers [5].

continuous scalability and colocation facilities. This is the most critical aspect in constructing data centres, because expansion capability and the handling of additional data or customers are necessary which may impact the architecture of data centre in long run if scalability is not taken into account. Any future change in the data centre which requires more space, devices or other technical aspect must be effectively managed without affecting the key existing data centre elements.

- Security: The data centre is required to store the client information and process online payment and the orders of their customers. The process which the data centre will be handling the most will be the payment process and the orders, in which the payment should be done in the most secure way. It is also critical that the proposed data centre has the highest data protection standards. In order to ensure both internal and external threats, the data security must include cyber security measures and physical security. The advantages of a high-security data centre ensure data integrity and keep trust between customers. As a data centre hosts information, software and facilities that companies use every day, organisations must ensure that they utilise adequate data centre protection. Lack of effective data centre protection may contribute to privacy abuse when confidential information regarding the business is leaked or compromised.

- Manageability: Manageability in data centres is about the responsibilities and processes associated with IT infrastructure management. A survey conducted among 300 data centre professionals by Tintri found that 49% of the respondents identified their biggest concern as manageability [6]. As manageability is such a critical part of the data centre, with authors in [7] categorising and evaluating fog data management, Meza has planned the data centre with manageability as one of its focus areas. Modern data centres rely on automation to improve manageability. Meza, with years of experience, has analysed the use case extensively and decided the data centre will be using a Data Centre Infrastructure Management (DCIM) software (**Figure 4**).

According to [9], DCIM covers monitoring, measuring, managing and controlling data centre utilisation and energy consumption of all IT-related equipment and facility infrastructure components. These equipment and components include power



Rooms	Racks	Devices
4	24	243

Total: 254

Not in Hierarchy	Connection Lost	Not Being Monitored
1	16	10

Used: 240U
Space Capacity: 1008U

80.31kW IT Equipment Power

Used: 80.30kW
Power Capacity: 274.20kW

- Unsupport: 106
- Unknown: 0
- Fault: 0
- Normal: 148

Name	Description
No records found.	

Name	Temperature°C
Hillsboro-DC > RM2 > RW > Rack 1 > ip-192-168-1-11.us-...	28.0°C
Hillsboro-DC > RM2 > RW > Rack 1 > ip-192-168-1-12.us-...	28.0°C
Hillsboro-DC > RM2 > RW > Rack 1 > ip-192-168-1-13.us-...	28.0°C
Hillsboro-DC > RM2 > RW > Rack 1 > ip-192-168-1-14.us-...	28.0°C

Room Name	Timestamp	Action
RM2	2019-4-4 11:54:00	Acknowledge

Figure 4.
A commercial DCIM software developed by Intel [8].

distribution units, servers and network switches to name a few. A typical data centre has a lot of workload. The workload increases immensely depending on the size of the data centre. For the food delivery company with millions of users, the data centre will have a huge workload if managed manually, and it would be unrealistic and impossible to run effectively. DCIM does tasks which are otherwise performed by data centre personnel. An important feature of DCIM is the real-time [10] central dashboard which displays information about critical systems from sensors and equipment. Data centre personnel are more informed about the operations and are likely to predict the next outage and avoid it. In addition to this, DCIM can handle non day-to-day tasks such as management of change. Therefore, DCIM is a critical piece of software for improving data centre manageability. The use of DCIM by Meza in this data centre will have huge benefits, resulting in less downtime and making manageability more robust (**Figure 5**).

Since cabling performance is a major factor in system outages, Meza will be using cabling from providers that ensure their cables can sustain higher performance. Meza hopes that using high-quality data centre fabric can reduce system outages and increase the overall manageability of the data centre.

- Cost: All business organisations strive to put the best performance with the lowest possible cost. It is the interest of both Meza and the food delivery company to bring the cost down while meeting business requirements. Total Cost of Ownership (TCO) is an estimate which includes the building the data centre and operating it. For the food delivery company, it is necessary that the TCO of building and operating a data centre is lower compared to hosting their application on a public cloud such as Amazon Web Services (AWS). According to [12], the largest driver of cost is determined to be the unnecessary unabsorbed costs

> resulting from the oversizing of the infrastructure. Meza has decided to deploy an adaptable physical infrastructure system. An adaptable physical infrastructure system reduces the waste due to oversizing substantially. As a result, the total cost of ownership is reduced too (**Figure 6**).

As shown in **Figure 6**, the room capacity design is non-adaptable (a), at the beginning, compared to adaptable physical infrastructure system (b), as the load increases.

In addition to this, Meza plans that with the use of Data Centre Infrastructure Management (DCIM) software, the operating costs can be reduced. One of the fundamental features of DCIM software is the use of automation across the board. Automation reduces manual labour with situational awareness. For example, resources such as energy can be increased during peak hours automatically instead of having maximum performance all day long regardless of the load. Moreover, use of DCIM software also allows data centre personnel to predict the life cycle of physical

cxtec
Are you getting the best performance, design and *manageability* in your data center?
"65% of system outages are related to cabling." –AFCOM: Data Center Institute/Gartner estimate
performance
design
manageability

Figure 5.
Cabling contributes to large number of system outages [11].

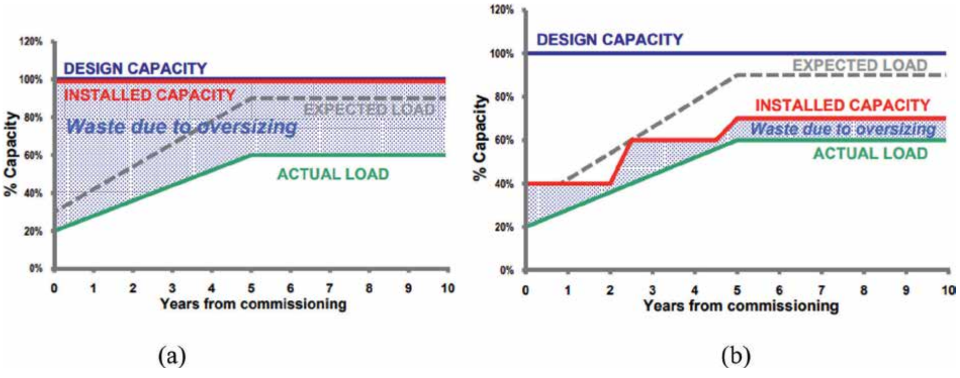


Figure 6.
Charts showing waste due to oversizing between non-adaptable (a) and adaptable (b) approaches [13].

infrastructure equipment, so they can change the equipment before they become faulty without compromising additional equipment due to failure.

2.3 Environment

The data centre will be located at Cyberjaya, and it is a specialised Information Technology district in Kuala Lumpur, Malaysia. The location is around 30 minutes away from the Kuala Lumpur city centre as well as the Kuala Lumpur International Airport.

Geographically, Malaysia is a well-known, stable region where natural disaster risk like tsunamis and earthquakes is extremely low. Especially in Cyberjaya, the district is above sea level throughout the year, so there is nearly no chance for massive flood happening. Cyberjaya is the core of Multimedia Super Corridor (MSC Malaysia), with MSC status, it guarantees the world-class infrastructure for IT industry and 99.9% guaranteed reliability in advance telecommunication technologies. The rental rate is from MYR 2.50 per square foot [14].

The environment is highly secured, it has a state-of-the-art CCTV system integrated with Malaysia's Emergency Response System and police personal monitoring the CCTV footage all the time with a quick response time for any emergency. These create a secure environment for the community (**Figures 7**–**9**).

Figure 7.
(NTT, ND).

Figure 8.
The environment of Cyberjaya [15].

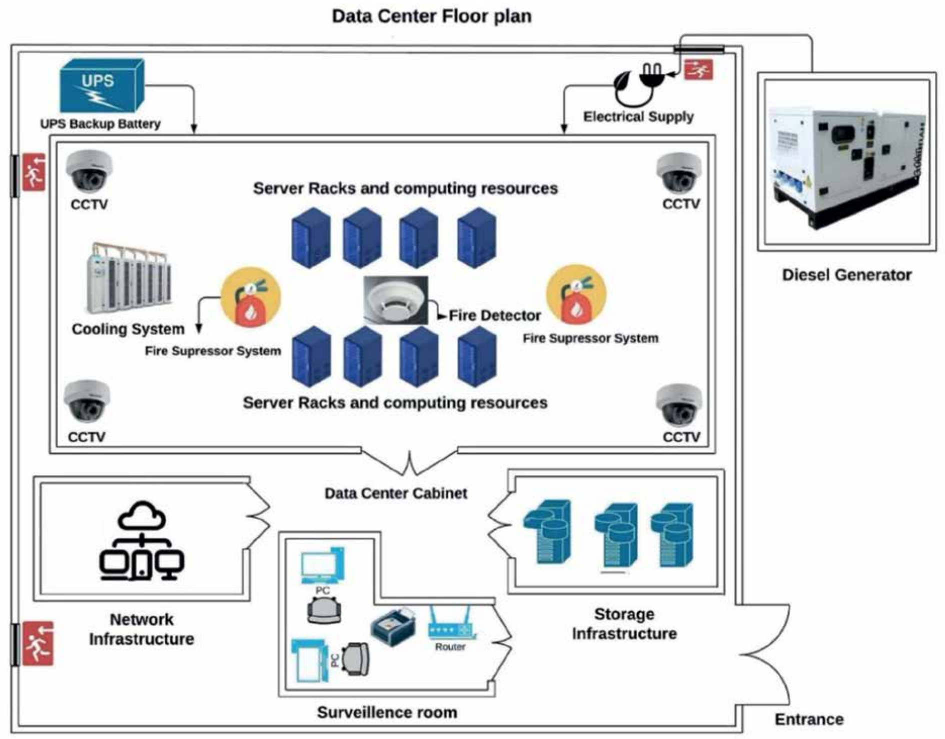


Figure 9.
Illustrate the proposed data centre floor plan.

3. Data centre design

3.1 Data centre floor plan

3.1.1 Floor plan justification

The above data centre floor plan design consists of 8 unique components that are necessary for a data centre including a surveillance room for monitoring the physical security and to create daily reports and analytics. The components that have been included are:

- Electrical supply: it is used to provide electricity for the entire data centre.

- Cooling system: it has been placed to provide cooling for the server racks and to avoid overheating when the computing resources are in use.

- Computing resources: they are responsible for handling all the processing powers for the database and to help with communications between the clients and the customers.

- Sever racks: it is used for the space allocation for the computing resources.

Figure 10.
Racks inside a data centre [17].

- Network infrastructure: it is responsible for a smooth communication between individuals as well as for faster and extra-secured payment [16].
- Storage infrastructure: this is where all the information is stored like details about the restaurant menu, customer credentials and so on.
- Fire detection system: in case of any overheating occurred by the computing resources or electricity overload, fire detection system will be able to detect it earlier and take necessary steps.
- Fire supressing system: in the event of any components catching fire, this system will be responsible for supressing it.

Furthermore, the data centre has been equipped with an Uninterruptible Power Supply (UPS) backup battery and a diesel generator in case of a power failure in order to achieve a higher availability for the system. Lastly, state-of-the art closed circuit television (CCTV) has also been placed in the data centre cabinet to be monitored remotely by higher officials (**Figure 10**).

4. Data centre components

4.1 Racks

In a data centre, racks can be considered as the building blocks. Traditionally, racks were mostly used for stacking IT equipment and saving floor space. However, racks in data centres today play a vital role in mounting heavy IT equipment, providing an organised environment for power distribution, air flow distribution for better cooling performance and cable management among many features [18]. Data centres demand a rack infrastructure that can mount a variety of equipment such as servers and switches. Therefore, it is important that the rack infrastructure can meet the requirements while offering sustainable performance.

4.1.1 Equipment in racks

The major equipment inside the rack will be the compute servers, storage servers and networking equipment such as switches. Different racks will have different compositions of these equipment.

- Compute servers

The main compute resources in a data centre are the servers. Most of the racks will be utilised for mounting rack servers for compute purposes. These servers are used for compute-intensive tasks such as processing and database hosting. These servers will be using enterprise-level processors such as Intel Xeon or AMD EPYC which have multiple physical cores providing high-level performance.

- Storage servers

Similar to compute rack servers, storage servers are mounted in the racks. Storage servers have a high density of storage capacity such as hard disks and SSDs. The emphasis on processing power in storage servers compared to compute servers is less. Therefore, storage servers typically use much less RAM and less performant processors. More on storage infrastructure will be discussed in this proposal.

- Switches

Switches act like a hub which connects different equipment such as servers in the rack with other servers or racks in the data centre. They are an integral part of the networking infrastructure.

4.1.2 Rack enclosures

Selecting a rack for a data centre requires consideration into some criteria such as dimension, design, capacity and material. According to [19], rack is available in three major types: open frame racks, rack enclosures or cabinets and wall-mount racks (**Figure 11**).

Rack enclosures or cabinets are a rack with four posts, doors and panels on the side. Depending on the design and manufacturer, the side panels can be removed to offer maximum flexibility. Among the most distinctive features of rack enclosures are airflow management, security, cable management and power distribution. These types of rack are ideal for use cases where the rack needs to store heavier equipment, hotter equipment and higher wattages per rack [19]. Doors that are on the front and back of the rack are ventilated for better airflow. Additionally, doors provide some levels of security. Most rack enclosures come with doors that can be locked which provide an additional layer of security (rack-level). Rack enclosures have a means of

Figure 11.
42 U rack enclosure or cabinet [20].

providing dedicated power distribution units (PDU) for the rack. The PUDs in rack enclosures are installed at the back or on the side, so they provide power without congesting the space inside the rack.

The size of the rack depends on many attributes. Some of these include:

- Width and depth of equipment used in rack

- Total weight of the IT and non-lT equipment (load rating)

- Number of cables entering the rack

- Rack units (RU) occupied

Most equipment used in racks are standardised with a width of 482.6 mm or 19inches. This current standard of 19-inch was established by Electronic Industry Alliance (EIA) [18]. In racks, the usable vertical space is measured in rack units. A rack unit is equal to 1.75 inches in height. Although racks consisting of deeper equipment and higher cable densities drive the need for a bigger rack size, the most widely used rack dimension is 42 U tall, 600 mm wide and 1070 mm deep.

Depending on the equipment mounted inside the rack, the rack can be considered a server rack or a networking rack. In comparison with server racks, network racks are much wider as they need additional room for cabling.

4.1.3 Justification

Based on the three types of racks, rack enclosures or cabinets will be used across the data centre. Since the data centre is going to be a newly built, wall-mount racks

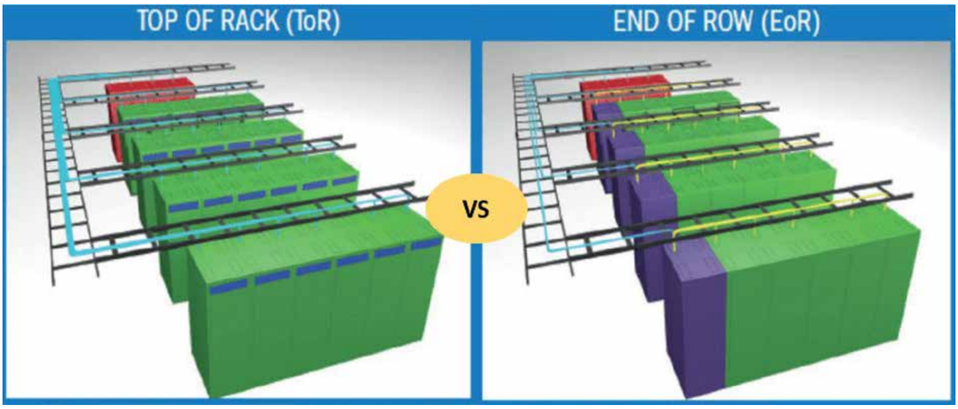


Figure 12.
Top-of-rack vs. end-of-row architecture [23].

can be avoided because there is enough floor space inside the data centre for the planned capacity. Compared to a wall-mount rack, the other two racks provide more racking of equipment for a given floor space. While open-frame racks offer a lot of features for a much lower cost than rack enclosures, the features such as better airflow control and better security are too important to be overlooked. Open-frame racks offer very little control over airflow. In addition to this, the use of side panels in rack enclosures prevents unrestricted hot air flowing inside the rack, heating up the equipment unnecessarily. According to [21], between 30 and 55% of a data centre's energy consumption goes into powering its cooling and ventilation systems. It is important that the racks chosen for the data centre can lower the cost of overall cooling as much as possible. In general, low cost racks such as open-frame racks have a significant effect on how much time it takes to complete rack-based work due to inefficiencies in areas such as cable management or mounting [18]. In [22], a decision support model has been proposed to the use of liquid-based cooling to measure and assess the waste heat resource accessible from retrofit within the High Performance Computing (HPC) and data centre (DC) industry (**Figure 12**).

As the data centre will be using top-of-rack switching, the use of networking racks will be limited. Top-of-rack switching architecture is considered for this data centre because it provides the benefit of better cabling, future-proofing with emerging standards and better support for multi-core servers by offering more bandwidth with low latency [24, 25]. Top-of-rack architecture avoids the number of cables going to the networking server considerably. Therefore, the size of racks in the data centre will be consistent.

Based on the consideration of attributes, the data centre will be using the standard 42 U tall, 600 mm wide and 1070 mm deep racks. Most servers mounted in the server will be 2 U. 2 U servers offer more advantages than a smaller 1 U server or oversized 5 U server. Due to limitation of physical size, 1 U server is affected by heating issues. While, 5 U servers are more powerful, they are more expensive and less cost-effective. Therefore, 2 U servers offer a compromise between performance and cooling [26]. When using standard equipment, the oversizing of data centre is not necessary. 42 U tall racks also provide several additional benefits [18]:

- Cheaper than taller racks

- No need for a ladder to reach all positions of the rack

- Less likely to interfere with overhead equipment such as fire suppression sprinklers

In conclusion, 42 U rack enclosures provide better features and are more suitable to be used in this data centre.

4.2 Storage infrastructure

In modern data centres, storage is becoming a highly complex component with increasing demands to store more and more data. Storage infrastructure for a data centre includes architectures, hardware equipment such as hard disks, SSDs and so on. Storage infrastructure in a data centre is tightly coupled with the networking for accessibility and delivery. In today's world, there are two challenges to high-performance storage systems: capacity and performance [27].

Capacity: Usage of computers, Internet of things (IoT) devices, mobile phones and other digital equipment has created a high demand for data storage. Data storage is increasing at a rapid pace every day. With the advancements in technologies such as image quality, average file sizes have risen considerably. As a data centre, the facility needs to have a storage infrastructure that has the capacity to meet these demands while offering the best performance possible.

Performance: Data centres need to focus on the storage performance regardless of the capacity requirements. It is vital for the storage infrastructure to be scalable and highly available. While storing hundreds of terabytes of data, unoptimised and poorly

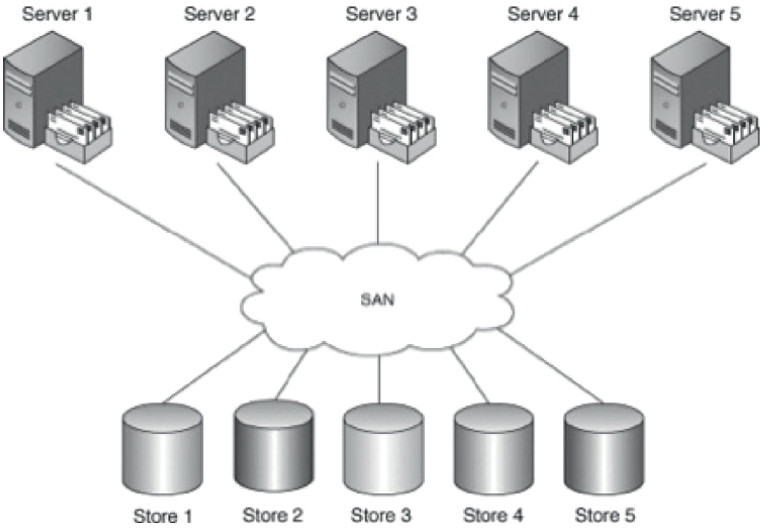


Figure 13.
Storage area network [28].

designed infrastructure could lower the performance of the overall data centre as data can be used in other areas like compute. The storage infrastructure in the data centre must be able to handle these requirements while overcoming the challenges faced. Traditionally, data centres use 3 popular storage solutions [27] (**Figure 13**).

4.2.1 Storage area network

Storage Area Network (SAN) is a dedicated network consisting of multiple storage devices. A SAN is a pool of block-level storage resources. SAN provides a higher level of management with the inclusion of multiple servers which manage data access and storage management [29]. Additionally, a SAN uses high-speed cabling and dedicated networking equipment such as switches. Modern SANs are based on fibre channel that can deliver high bandwidth and throughput with data speeds of up to 16GB per second. With the reductions in Solid State Drives (SSDs), SAN can consist of SSD arrays which offer much more I/O performance than Hard Disk Drives (HDDs). Although SAN is complex to deploy and manage, it is highly scalable and available. Since SAN runs on its own dedicated network, it does not face the issue of network-attached storage (NAS) solutions due to shared bandwidth and network congestion (**Figure 14**).

A SAN consists of various components which can be grouped into 3 main categories [30]. These categories are Host components, Fabric components and Storage components.

- Host components

These components are located in the computer servers or any other type of server accessing the SAN. Compute servers (hosts) use a host-based adapter (HBA) which has a fabric port that enables communication between the server (host) and SAN switches.

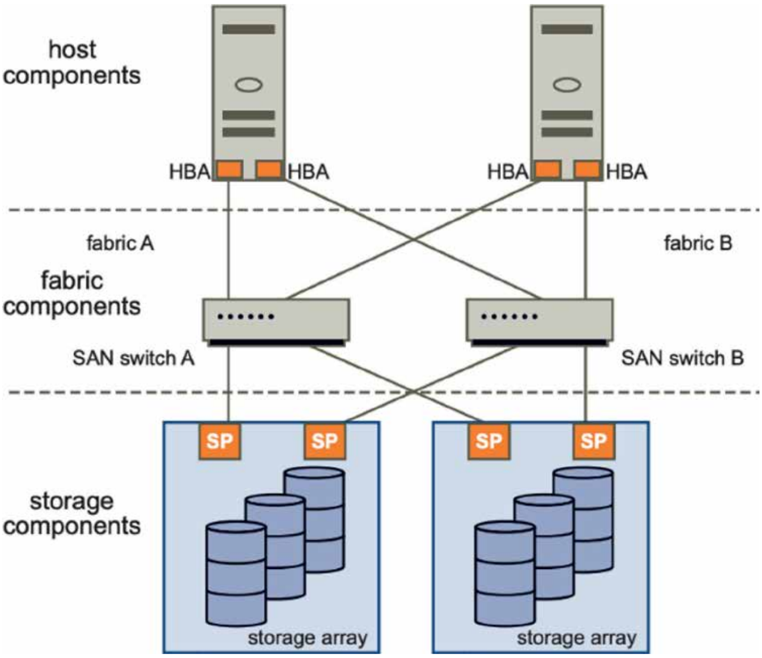


Figure 14.
SAN component layers [30].

- Fabric components

Fabric components include the switches, cables and communication protocols [30]. The switches used in SAN according to the SAN topology will be Fibre Channel (FC) switches. These switches will provide 64-128 ports per switch and have built-in fault tolerance. Since this SAN uses FC, the majority of the cables used in the SAN would be fibre optical cables. Fibre optical cables provide higher bandwidth and data speeds. In addition, the fabric components define the communication protocol. For this SAN, FC is used as the protocol, and based on that, a switched fabric topology is used.

- Storage components

The fundamental parts of any SAN are the storage components. Storage components are the storage arrays. Storage arrays contain storage processors which communicate with disk arrays. In this proposed data centre's storage infrastructure, the SAN will use SSD disk arrays. SSDs are one of the fastest storage mediums available today (**Figure 15**).

The SAN will use Core-Edge topology which is based on the switched fibre channel. The two most important traits of Core-Edge topology are the resiliency and performance that this topology provides. In this topology, two or more core switches are used to interconnect two or more edge switches. Edge switches can be the switches that connect with core switches from servers or disk arrays. In addition to this, the use of this topology in SAN will encourage a balance between usable ports and dedicated inter-switch communication [31].

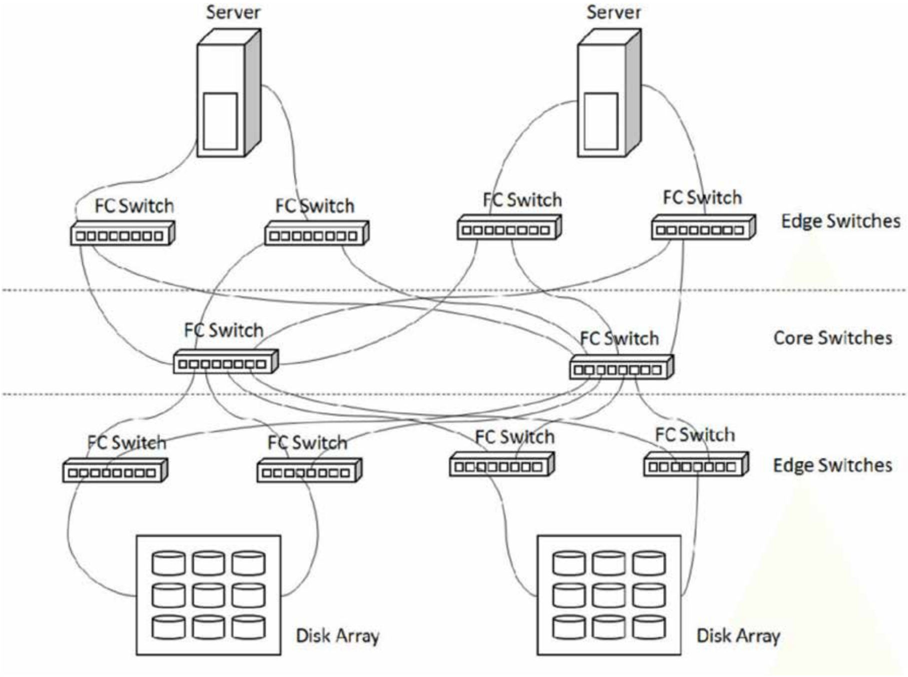


Figure 15.
Core-edge SAN topology [31].

4.2.2 Justification

Based on the comparisons made above, Meza's new data centre will be using a Storage Area Network (SAN). The growing food delivery company's active users are increasing, and they require a scalable storage solution. Therefore, DAS with no scalability cannot be chosen. While NAS is cheaper and easier to maintain, SAN offers better performance. For a large organisation and data centre, SAN is ideal. Another key factor is that SAN works with virtualisation [32]. Virtualisation is a popular technology not only used in data centres but also heavily used even today. Other benefits of SAN include improved storage utilisation, better data protection and recovery and elimination of network bottlenecks [33].

A key difference in how data is stored is that SAN uses block-level storage, while NAS uses file-level storage. The biggest advantage of block-level storage is that it offers better access and control privileges. This is critically important since the food delivery company already has 5 million users, and easier management of users' files is a key business requirement.

As for the SAN technology, Meza will be choosing the Fibre Channel (FC). The key factor in making this decision is that FC provides significantly better performance and reliability. For such a scale of growing 5 million active user base, performance and reliability are crucial. It is possible to build a storage network of thousands of nodes without affecting throughput and latency. In addition to this, the SAN will use arrays of SSDs instead of HDDs (**Figure 16**). SSDs provide significant increase in speeds, and the price difference between the two has narrowed over the past few years [34].

The topology for SAN infrastructure used is Core-Edge. According to [35], SAN designs should always use two isolated fabrics for high availability. Since this data centre is a tier 4 data centre, high availability and resiliency are crucial. One of the reasons why Core-Edge FC is selected is because point-to-point or FC-AL does not have high availability, as in if case one link fails, the entire storage network becomes

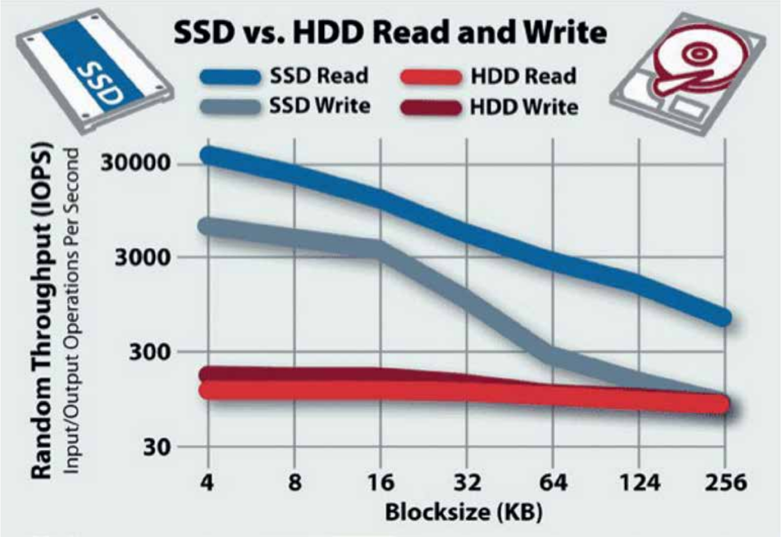


Figure 16.
SSDs have higher read and write speeds over HDDs [34].

unavailable. Finally, Core-Edge supports millions of nodes, offering high level of scalability [31]. Scalability in storage is imperative as it needs to continuously grow every day.

5. Conclusion

We can conclude by saying that the world of IT is constantly increasing, and this will never stop the demand for innovative and better solutions. There is no question that future confirmation of the solution and equipment chosen for this task is same. From security to smart execution, the planned data centre is carefully considered. Scalability, CO_2 reduction, system resilience, sustainability, applying machine learning and other emerging technologies are important to be considered for the data centre design. Moreover, the colocation system allows clients to locate their data by renting a space in the data centre and choosing the equipment. The focus on this role should be incredibly strong if carried out and can make sure every requirement of the food ordering system is encountered.

Conflict of interest

The authors declare no conflict of interest.

Author details

Yaseein Soubhi Hussein[1*], Maen Alrashd[2],Ahmed Saeed Alabed[1] and Saleh Alomar[2]

1 Computer Science and Information Systems Department, Ahmed Bin Mohammed Military College, Qatar

2 Faculty of Science and Information Technology, Jadara University, Irbid, Jordan

*Address all correspondence to: dr.yaseein@abmmc.edu.qa

References

[1] Colocation America. Data Center Standards (Tiers I-IV). 2015. Available from: https://www.colocationamerica.com/data-center/tier-standards-overview.htm

[2] CtrlS. Significance of Tier 4 Data Center. 2014. Available from: https://www.ctrls.in/blog/significance-tier-4-data-center/

[3] Greengard S. Data Center Tiers: Formulating a Strategy. 2019. Available from: https://www.datamation.com/data-center/data-center-tiers.html

[4] Impact. Tier IV Data Centers. 2009. Available from: https://www.impactmybiz.com/blog/blog-why-you-need-a-tier-iv-4-data-center/

[5] WHOA.com. Tier IV Data Centers. 2017. Available from: https://www.whoa.com/data-centers/

[6] DCNewsAsia. Manageability Top Concern for Data Center Professionals. 2016. Available from: https://datacenternews.asia/story/manageability-top-concern-data-center-professionals

[7] Sadri AA, Rahmani AM, Saberikamarposhti M, Hosseinzadeh M. Fog data management: A vision, challenges, and future directions. Journal of Network and Computer Applications. 2021;**174**:1-24

[8] Intel. Intel® Data Center Manager. 2020. Available from: https://www.intel.com/content/www/us/en/software/intel-dcm-product-detail.html

[9] Gartner. Data Center Infrastructure Management (DCIM). 2020. Available from: https://www.gartner.com/en/information-technology/glossary/data-center-infrastructure-management-dcim

[10] Javadzadeh G, Rahmani AM, Kamarposhti MS. Mathematical model for the scheduling of real-time applications in IoT using dew computing. The Journal of Supercomputing. 2022;**78**:7464-7488

[11] CXtec. Just How Manageable is Your Data Center?. 2020. Available from: https://www.cxtec.com/resources/blog/just-how-manageable-is-your-data-center/

[12] Rasmussen N. Determining Total Cost of Ownership for Data Center and Network Room Infrastructure. 2015. Available from: https://download.schneider-electric.com/files?p_File_Name=CMRP-5T9PQG_R4_EN.pdf

[13] Rasmussen N. Avoiding Costs from Oversizing Data Center and Network Room Infrastructure. 2015. Available from: https://download.schneider-electric.com/files?p_File_Name=SADE-5TNNEP_R7_EN.pdf

[14] Malaysia C. Available from: https://www.cyberjayamalaysia.com.my/community/overview

[15] Richard. Essential Information about Cyberjaya - Malaysia's Technology and Innovation Hub. 2019

[16] Hussein Y, Alrashdan M. Secure payment with QR technology on university campus. Journal of Computer Science & Computational Mathematics. 2022;**12**:31-34

[17] Facebook. Opening our Newest Data Center in Los Lunas, New Mexico. 2019. Available from: https://engineering.fb.com/data-center-engineering/los-lunas-data-center/

[18] Pearl H, Wei Z. How to Choose an IT Rack. 2015. Available from: https://

download.schneider-electric.com/files?p_Doc_Ref=SPD_VAVR-9G4MYQ_EN

[19] Tripp Lite. Rack Basics: Everything You Need to Know Before You Equip Your Data Center. 2018. Available from: https://www.anixter.com/content/dam/Suppliers/Tripp%20Lite/White%20Papers/Rack-Basics-White-Paper-EN.pdf

[20] Tripp Lite. 42U SmartRack Standard-Depth Rack Enclosure Cabinet with Doors, Side Panels & Shock Pallet Shipping. 2020. Available from: https://www.tripplite.com/42u-smartrack-standard-depth-rack-enclosure-cabinet-doors-side-panels-shock-pallet-shipping~SR42UBSP1

[21] DataSpan. Data Center Cooling Costs. 2019. Available from: https://www.dataspan.com/blog/data-center-cooling-costs/

[22] Ljungdahl V, Jradi M, Veje C. A decision support model for waste heat recovery systems design in data Center and high-performance computing clusters utilizing liquid cooling and phase change materials. Applied Thermal Engineering. 2022;**201**:1-10

[23] Parés C. Top of the Rack vs End of The Row. 2019. Available from: https://blogs.salleurl.edu/en/top-rack-vs-end-row

[24] Juniper Networks. Next Steps Toward 10 Gigabit Ethernet Top-of-Rack Networking. 2016. Available: https://www.juniper.net/us/en/local/pdf/whitepapers/2000508-en.pdf

[25] Hussein YS. Impact of applying channel estimation with different levels of DC-bias on the performance of visible light communication. Journal of Optoelectronics Laser. 2021;**40**

[26] Thinkmate. 2U Rack Server. 2017. Available from: https://www.thinkmate.com/inside/articles/2u-rack-server

[27] Scala Storage. Scala Storage Scale-Out Clustered Storage White Paper. 2018. Available from: http://www.scalastorage.com/pdf/White_Paper.pdf

[28] Lee G. Storage Network. 2014. Available from: https://www.sciencedirect.com/topics/computer-science/storage-network

[29] RedHat. What is Network-Attached Storage?. 2020. Available from: https://www.redhat.com/en/topics/data-storage/network-attached-storage

[30] VMware. SAN Conceptual and Design Basics. 2016. Available from: https://www.vmware.com/pdf/esx_san_cfg_technote.pdf

[31] Gençay E. Configuration Checking and Design Optimization of Storage Area Networks. 2009. Available from: https://www.researchgate.net/publication/314245428_Configuration_Checking_and_Design_Optimization_of_Storage_Area_Networks

[32] Bauer R. What's the Diff: NAS vs SAN. 2018. Available from: https://www.backblaze.com/blog/whats-the-diff-nas-vs-san/

[33] Robb D. Storage Area Networks in the Enterprise. 2018. Available from: https://www.enterprisestorageforum.com/storage-networking/storage-area-networks-in-the-enterprise.html

[34] Rubens P. SSD vs. HDD Speed. 2019. Available from: https://www.enterprisestorageforum.com/storage-hardware/ssd-vs-hdd-speed.html

[35] Singh S. Core-Edge and Collapse-Core SAN Topologies. 2017. Available from: https://community.cisco.com/t5/data-center-documents/core-edge-and-collapse-core-san-topologies/ta-p/3149001

Chapter 2

Data Centre Infrastructure: Power Efficiency and Protection

Yaseein Soubhi Hussein, Maen Alrashd, Ahmed Saeed Alabed and Amjed Zraiqat

Abstract

The rapid expansion of e-commerce necessitates expanding the capacity and dependability of data centres in order to provide services with the proper level of quality. A green technology that has a lot of potentials to reduce CO_2 emissions is optimization data centre design. However, a large data centre required a large amount of electricity because the capacity of the racks is higher, which required more potent cooling systems, power supplies, protection and security. These will increase the cost of the data centre and render it unusable for services. In this chapter, we provide a design for a tire-four data centre that will be situated in Cyberjaya, one of Malaysia's best locations. This design's primary goal is to offer highly functional and high-quality e-commerce services, particularly food delivery. Each component of the data centre has been carefully developed to deliver a range of services, including the administration of IT infrastructure, co-location, cooling system and protection. Additionally, advice and support have been given to guarantee that the suggested design outperforms competing data centres in terms of dependability, power efficiency and storage capacity. The analysis, synthesis, and evaluation of each element of the proposed data centre will be considered in this chapter.

Keywords: data centre, power efficiency, power usage effectiveness (PUE), protection, cooling system, scalability

1. Introduction

Meza is one of the native data centre firms that offer a range of services, including the administration of IT infrastructure and top-notch security and data management. Meza is anticipated to be constructed in a number of data centres throughout Malaysia. More than five million people in Malaysia use a food delivery application company's services, and that number keeps growing. Due to the current infrastructure's inability to handle the massive quantity of data processing, users may have a bad user experience as a result of the lengthy response times from servers and delayed processes. In order to accommodate the company's continuous growth, Meza has been chosen by the corporation to build a data centre. Customer relationship management, which unifies all client communication into one inbox, will be needed by the data centre to conduct online food orders and payments.

For this chapter, a data centre design with the necessary components will be provided, and each component of the proposed data centre will be examined, synthesised, and evaluated. These data centre components including power usage effectiveness and efficiency, cooling system and protection. Other data centre components, such as storage infrastructure, networking and environment have been discussed in another chapter Data Centre Infrastructure Design and Performance.

2. Power system

2.1 Electrical power system

Power is an important element that gives life to a data centre and maintains the IT infrastructure even when an interruption takes place. The value of the power system in the data centre cannot be emphasised enough. Power is one of the significant factors when it comes to cost estimation of colocation services as well. Additionally, the type of currents used to power the servers, switches, routers and associated IT infrastructures are of two different types. One of them is DC (Direct Current) and the other is AC (Alternating Current) [1].

Meza has decided to use 277/480 (277 for Single phase or 480 for three-phase power) Volt AC power supply for the food delivery data centre because it removes the PDU (Power Distribution Unit) and passes directly the power to the server cabinet at a higher voltage. It is also energy efficient and provides a decreased load for the cooling systems. Furthermore, it has increased consistency [2]. Moreover, in [3] reinforcement learning method is applied for automating energy efficiency.

Since Tier 4 has been chosen for this data centre, there are a few components that will be discussed which will allow the data centre to be up and running 24/7. The components that Meza has decided to use are:

- Automatic Transfer Switch
- Backup Power Supply
- RPDUs (Rack Power Distribution Units)
- Backup Generator

2.1.1 Automatic transfer switch

When there is an interruption of power supply, an Automatic Transfer Switch (ATS) is used. It is an electrical switch that moves the source of the power supply from the main source to a backup source. If the primary power source detects a power loss, the ATS triggers the substitute power sources which will provide a continuous power supply [4]. There are four types of ATS, and after a copious amount of research, Meza has decided to use Closed Transition ATS. This switch functions in a way where a little pause is not accepted in the power source supply which allows it to detect a power blackout and power fluctuation earlier and allows the data centre to have a smooth switchover while running the main power supply and the backup supply simultaneously without any interruption of services in the data centre [4] (**Figure 1**).

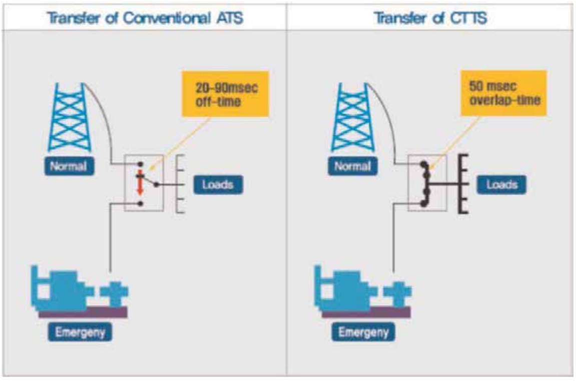


Figure 1.
A comparison between traditional ATS and closed transition ATS [5].

2.1.2 Backup power sources

The need for backup power resources is very crucial to provide with uninterruptable power supply for the data centre in the event that the primary source of power supply fails. Two of the backup power sources chosen are:

- Uninterruptible Power Supply (UPS): UPS backup power is a critical component. Without it, fluctuations in power and outages can push workloads down, damage equipment and result in hefty payments. The main aim of it is to preserve the data centre infrastructure until stable power returns or long-term alternative power backup systems are up. There are mainly three types of UPS, they are:
    - Standby/offline UPS
    - Line Interactive
    - Online/Double Conversion
- The Online/Double Conversion UPS has been recommended by Meza. Online/Double Conversion UPS is the kind of UPS that, during normal operations, provides AC load power through a Rectifier and Inverter Combo and uses an AC power inverter during power failure. The output power supply therefore still stays ON and switching is not required [6]. This UPS has been chosen because as stated by [7] that it provides the highest degree of security for a facility by shielding IT equipment from the raw power utility. This system operates by power conversion from AC to DC and back into AC. Additionally, this UPS is said to be the only kind of UPS with zero battery transmission time and is suitable for sensitive applications. That is why this UPS is the safest to use for any mission-critical equipment for any facility such as the food application data centre.
- Backup Diesel Generator: When the main power source is disrupted, backup generators are used to provide electricity for the data centre. Power interruption due to grid failures can lead to a high risk of operational loss in data centres. Furthermore, the system and components of data centres should function constantly and require a reliable, continuous power supply 24 hours a day, 7 days

a week. If the event of a power lapse, files can be lost or corrupted, mainframes can malfunction [8].

2.1.3 RPDUs (rack power distribution units)

The RPDU generates no power but rather delivers power from the available power supply. In the world of data centres, the RPDU is able to monitor, manage and regulate the power usage of many devices. It can supply vast quantities of electricity and can be accessed via the local or remote network. RPDUs can withstand high power density and are immune to greater temperatures in order to satisfy the ever-changing needs of the data centre [9]. Rack PDUs can be categorised into two categories which are non-intelligent PDUs and smart PDUs [10]. There are mainly four types of PDUs, they are:

- Basic Rack PDU
- Metered Rack PDU
- Monitored Rack PDU
- Switched Rack PDU

Since the aim of the proposed data centre is to be robust, highly reliable and highly effective, only a smart PDU will be discussed in this paper which will be the modern Switched Rack PDU.

Switched Rack PDU is a power management unit that can be installed on a typical industry rack and have the capability to power on and off remotely for individual outputs. Moreover, the Switched Rack PDU offers outlet control for rebooting locked devices as well as remote access capabilities to power and environment information. It also offers current, voltage, power (kW), apparent energy and cumulative energy measurements per outlet (**Figure 2**) [11].

Figure 2.
A switched rack PDU design [12].

2.1.4 IT space allocation

The above picture showcases an electrical power system supply arrangement in a data centre as well as how the cablings are handled (**Figure 3**).

2.1.5 Justification

As the data centre for the food delivery company has been decided to go with tier 4 data infrastructure, the electrical power supply has been tailored to those needs. Two main utility power grids have been arranged. The data centre has also been equipped with a Closed Transition ATS because this automatic transfer switch will make sure that there will not be any power outage when transferring the power from the main supply to the backup supply which will ensure a smooth service for the food delivery application. Furthermore, the data centre has been equipped with Online double conversion UPS because, with this, there is very less chance of electrical load loss. It is also efficient and has a good PUE (Power Usage Efficiency) as supported by [13]. Moreover, a backup generator has been placed in case both the power grid somehow blackouts.

Additionally, for the rack power distribution unit, Switched Rack PDU has been placed because it will allow the higher officials of the data centre to remotely monitor all the data load and electricity consumption and as well as remotely change the voltages for the racks/servers which will make the energy consumption of the data centre even less. This statement is also approved by [14]. Lastly, the power utilisation of this data centre is highly efficient because of the necessary features that have been added which are also supported by [15] that by adding features such as power-saving “standby” modes, energy management software and efficient cooling systems, data centres can become more energy-intensive. Such improvement in efficiency will produce significant energy savings and reduce the electricity grid load.

2.2 Fire detection system

Fire detection systems are made for the early identification of fires while time for the safe evacuation of individuals is still available. In order to ensure the health of

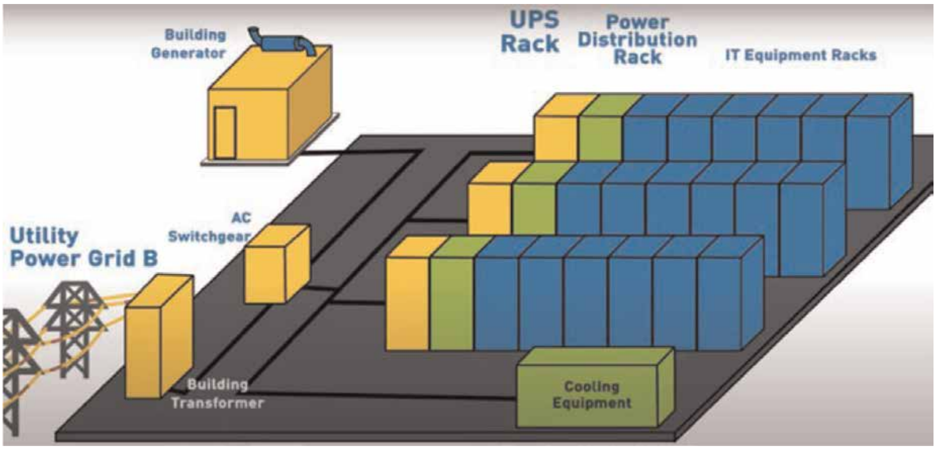


Figure 3.
Proposed IT space for the electrical power supply [10].

emergency response workers, early detection also plays an important role. Furthermore, fire damage and operational downtime can be minimised as monitoring measures begins while the fire remains low. Most alarm systems supply emergency responders with information about the location of the place where the fire has started which speeds up the process of controlling the fire [16].

There are mainly three levels in a data centre that needs to be protected, they are:

- Building Level: The primary goal is to avoid fire in the building and its workers. Fire sprinklers and portable extinguishers are the most common forms of fire safety.

- Room Level: This level focuses on specific rooms such as the data centre cabinet, surveillance room, etc.

- Rack Level: The last standard of fire safety at a data centre is on the rack level. Data servers, storage and network infrastructure must be protected using this fire protection to minimise fire damage (**Figure 4**) [17].

So, to minimise the downtime and loss of data for the food delivery data centre in case of a fire breakout, the recommended smoke detector, and the fire alarm system will be discussed in this chapter.

2.2.1 Smoke detector

A Smoke Detector is an electronic fire detection unit which senses the presence of smoke automatically. Smoke detectors are typically managed by a central fire alarm device, operated by building power with a battery backup in building infrastructures [18]. There are typically 3 types of smoke detectors. Air-aspirating/air-sampling smoke detector has been recommended by Meza.

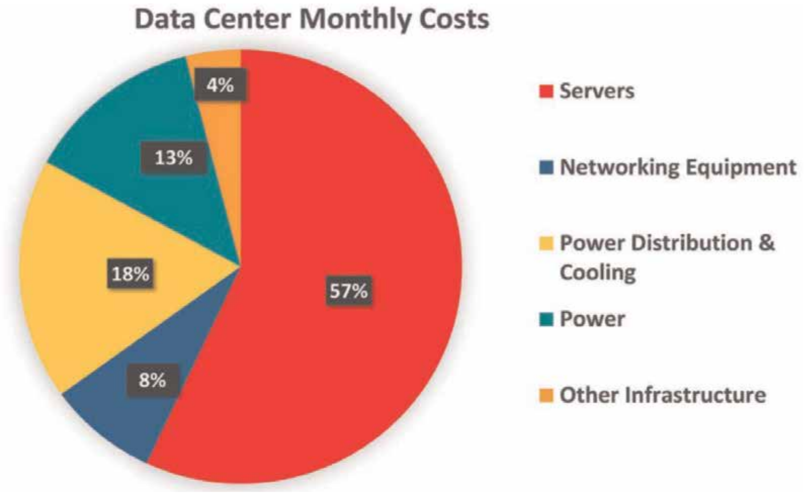


Figure 4.
The importance of rack-level protection by illustrating a monthly cost [17].

2.2.1.1 Air-aspirating/air-sampling smoke detector (ASD)

This smoke detector has become very popular as it can detect fires at a very early point, even before smoking happens as well as before an open fire and before excessive smoke happens. ASD can locate fires considerably faster than point or beam detectors, meaning the first signs of smoke can be reacted to quickly. This early detection is important for sensitive and high-risk infrastructure. These detectors can also be set to a traditional point detector sensitivity level (**Figures 5** and **6**) [21].

2.2.2 Fire alarm system

Once the smoke has been detected, the smoke detector will notify the fire alarm system. A Fire Alarm System is intended to alert people to an emergency in order to protect themselves. Whatever detection system it is, the sounders can be used to alert building staff about the risk of fire or evacuation if an alarm is activated. The Fire Alarm Control Panel is the 'brain' of the fire detector system. The central hub gives a status indicator to users for all the detector signals. This system can also be programmed to simulate an alarm for regular fire drilling and evacuation, so that all workers know what to do in case of a real fire [22].

Generally, fire detectors are of three types which are Conventional fire alarm system, Addressable fire alarm system and Wireless fire alarm system as stated by [23]. After plentiful amount of research, it is believed to go with Wireless Fire Alarm System for the food data centre infrastructure because the data centre already has many cablings with the IT equipment and resources. So, wireless fire alarm system has been chosen over addressable fire alarm system and regarding the cost efficiency, they both will cost roughly similar price because even though wireless alarm system is expensive but for an addressable fire system, it will take a high cost to arrange all the cablings as stated by [24].

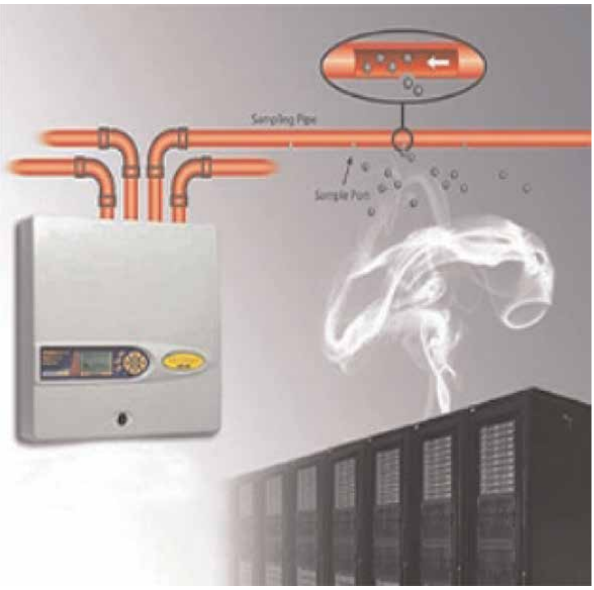
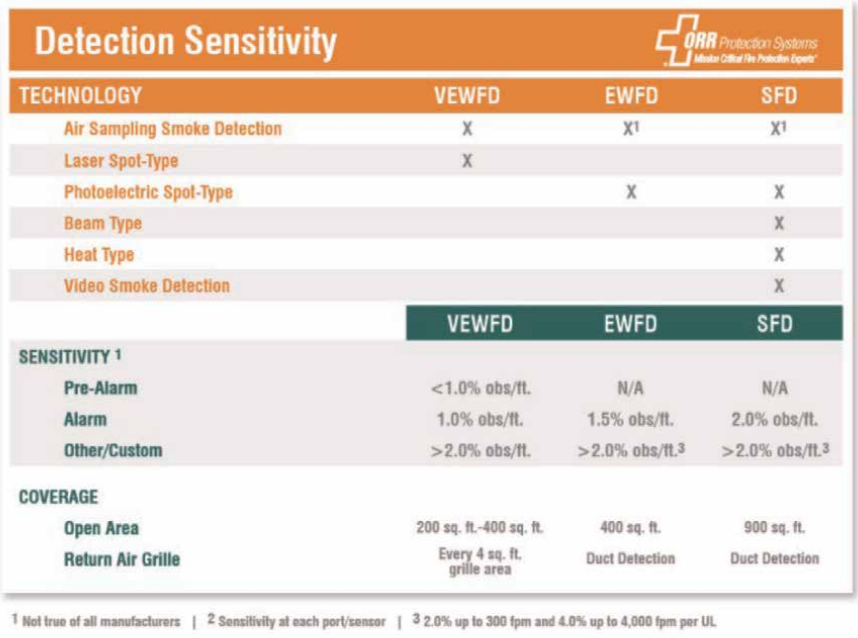


Figure 5.
How an air-sampling smoke detector works [19].



Detection Sensitivity

TECHNOLOGY	VEWFD	EWFD	SFD
Air Sampling Smoke Detection	X	X1	X1
Laser Spot-Type	X		
Photoelectric Spot-Type		X	X
Beam Type			X
Heat Type			X
Video Smoke Detection			X

	VEWFD	EWFD	SFD
SENSITIVITY 1			
Pre-Alarm	<1.0% obs/ft.	N/A	N/A
Alarm	1.0% obs/ft.	1.5% obs/ft.	2.0% obs/ft.
Other/Custom	>2.0% obs/ft.	>2.0% obs/ft.3	>2.0% obs/ft.3
COVERAGE			
Open Area	200 sq. ft.-400 sq. ft.	400 sq. ft.	900 sq. ft.
Return Air Grille	Every 4 sq. ft. grille area	Duct Detection	Duct Detection

1 Not true of all manufacturers | 2 Sensitivity at each port/sensor | 3 2.0% up to 300 fpm and 4.0% up to 4,000 fpm per UL

Figure 6.
Performance levels for fire detection systems [20].

Wireless Fire Alarm System: The solution of choice for many applications is wireless fire alarm device. The huge versatility and endless combination of wireless alarm devices make it a good choice for organisational sensitivity. Each unit of the range communicates with self-optimising amplitude and frequency through sophisticated bidirectional encrypted radio transmission. Additionally, multi-directional integrated antennas guarantee the virtual removal of signal corruption. Furthermore, wireless fire alarm systems have shown that they provide the best security in the premises in a reliable and cost-effective manner (**Figure** 7) [26].

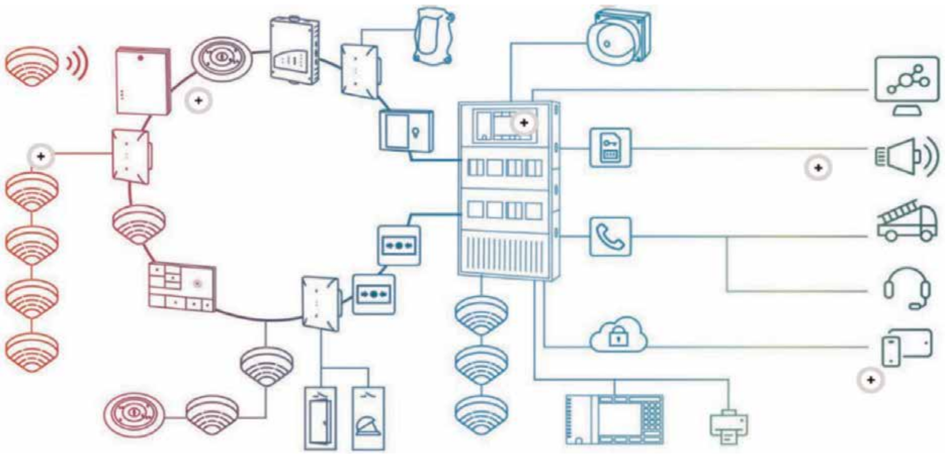


Figure 7.
How a fire detection system along with a fire alarm system generally operates [25].

2.2.3 Recommendation

After critically analysing the different types of smoke/fire detectors, Meza has decided to use Air-Sampling smoke detector because since the data centre will handle a huge load and huge processing power for the food delivery infrastructure, this early smoke detection system should be installed in case there is an overload and overheating which may cause a fire. As supported by [20], it can help with relatively early fire warning detection in rooms containing IT and telecommunications equipment. Additionally, the very early warning of smoke for the staff running the infrastructure is a crucial aspect of air sampling smoke detectors. The early warning capability enables managers to evaluate a smoke long before it enters an emergency state and activates a fire suppression system.

Lastly, wireless fire alarm system has been suggested as well because the data centre needs to be protected explicitly from fires and needs to be cost-effective and more reliable. As stated by [27], wireless fire alarm system is easier to deploy with low downtimes and has easier maintenance. It also has higher reliability, is more cost-valuable in the long run and can be repositioned easily if necessary which are the key points that are being kept in mind by Meza to make the data centre more efficient, robust and safe; see **Figure 6**.

3. Fire suppression systems

For a data centre fire suppression system is mandatory. The food ordering application holds many data and computes many processes and it is mandatory to keep the data centre safe from a fire outbreak if happens. A fire suppression system is a collection of designed units installed with the application of a material to extinct flames. The fire protection device typically has a built-in component that identifies fires by flame, smoke and other alert signals at the beginning stages. These are connected to an alarm device to warn in the presence of the fire and to take action to avoid the fire. Most fire detectors activate an application of an exterior material upon identification and/or warning immediately to extinguish the fire. However, certain fire suppression devices are issued manually (**Figure 8**).

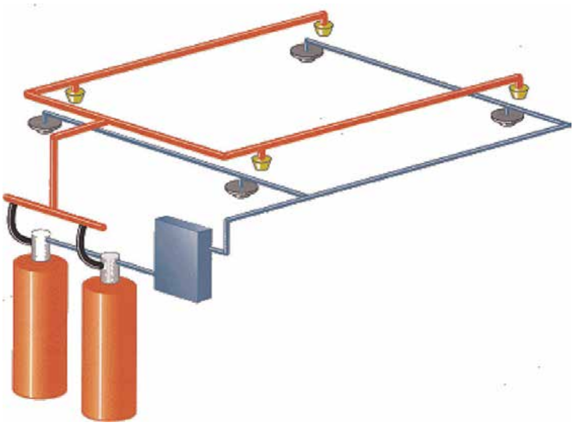


Figure 8.
Fire suppression system 1 [28].

3.1 Level of protection

There are three levels of protection for data centre. Which will ensure the safety of the data centre and the information stored in the system. First level of protection is building-level fire protection. The key goal is to defend the buildings and their workers from fire. Fire sprinklers and handheld extinguishers are the most widely used type of fire protection. The construction policy for handheld extinguishers requires that, if the class (A) combustible materials are in the workplace, there is a portable fire extinguisher for every 3000 square meters. A building may also use passive fire safety, including the construction of firewalls and floor mounts which dramatically delay the expansion of the fire in other areas of the building [29].

Second level of protection is room-level fire protection. The National Fire Protection Association (NFPA) sets the standards for room-level protection. The water still occurs in the piping of a wet piping network which will escape instantly after triggering of the warning. The downside to this device is that the pipe will leak and spill on the room facilities. The most commonly employed space fire safety is a pre-action device. The triggering of the sprinkler device needs at least two fire detection points. Other devices divide the space into sprinklers, so they just go off in the quadrant triggered. The best solution for data centre fire safety is fire sprinkler devices. Novec 1230 and FM-200 are the two rising safe agent gas systems. By raising the fire heat by absorption, they contain the fire. Such gases have zero ozone loss, rendering them biologically and humanly free. The physical footprint is smaller than inert gas systems as no agent is required to occupy a whole space. Electrically non-conductive, non-corrosive, sterile agent gases do not leave any traces upon evaporation. It renders them the best fire prevention contractor in data centres. As with fire sprinklers, the space is equipped with a tube network [29].

Third level of protection is rack-level fire protection. Relevant appliances and loss control need this fire protection. Although the mandatory fire sprinters protect the building and the room from fire, the appliance is not unprotected, which is worth 57 per cent of the cost in the room. To order to conserve money, the hardware has to be secured from a fire on a shelf. The implementation of a preconceived automated fire deletion mechanism safeguards the unit with the identification and deletion of the fire within seconds until it is triggered by the complete flute or sprinkler machine. It avoids the disruption to the facilities done by a water-based sprinkler and allows huge amounts of agents to be released into an expensive overall flood container [29].

To ensure the protection of the data stored in the data centre all three levels of protection should be taken. This will help the Meza team to protect the data centre of the food ordering application system. And reduce the cost of damage.

3.2 Types of protection

The data centre for food ordering applications can be protected from fire in a variety of ways. The first is the water-powered sprinkler device that manages the flames, stops them from spreading and avoids structural harm. The sprinkler device is considered as an inexpensive and simple approach utilising about 25 gallons per minute of water. This leads to certain risks, which may be greater than the fire loss if electrical conductivity of data centre appliances, is because the spot where the fire takes place would tremendously wash, which will take the business more.

The Water Neck System, a recent entrant to the water-based fire protection program, is another water-based method. The spreading of the water in the specified area needs strong pressure pumps, which is advised only for wide areas, as it resulted in poor fire output in order to prevent flooding the area entirely, particularly in the fire is blocked. In fact, after suppression, the mist systems leave residual vapour, and costs and a problem exist with the equipment.

The clean agent, which can extinguish the fire very easily and connect the fire damage to the data centre equipment in its location which does not need water, is also a way of protecting the data centre from fire accidents. The main purpose of this sort is to protect the resources that are important, dynamic and essential. It is distinct from other forms as it does not involve extinction washing so no corrosive or residue is left behind. In complex environments, it can extinguish fires in blocked or three-dimensional areas. In addition, two forms of cleaning agents are usable. Firstly, halocarbon agents including carbon, hydrogen and halogens including fluorine, thus creating dangerous effects for those in the fire because fuel is being polluted [30] and causes a breathing problem. Secondly, inert gas agents are prepared for gasses such as nitrogen, argon and carbon dioxide which are less risky to humans and less damaging to the resources than the first type. Both agents are regarded as electrically non-conductive and can be used in typically covered areas.

3.3 Recommendation

Kidde, fire system, a global pioneer in the development and manufacturing of fire detection and safety systems between clean agents-inert gas type and water sprinkler, was the focus of the previous debate and several real-world experiments, such as a study. As a Meza team, it suggests the clean agent-inert gas network that is installed for the food ordering application data centre, as well as the different benefits and functionality it provides, so that a fire will easily be extinguished and damage in a certain region limited, as well as no cleaning after the fire takes place. Therefore, the expense is inexpensive to introduce and eventually it can be built and managed very effectively and does not affect the safety or atmosphere of people. Therefore, beyond the core of the data centre, we do need to use the sprinkler network for assisting in certain circumstances by positioning several halocarbon agents.

3.4 Network infrastructure

See (**Figure 9**).

3.4.1 Cabling

Despite too much emphasis on various forms of technology and how organisations really bring their networks to use, it’s simple to think about the physical framework that allows every data centre networking system feasible. Cabling is a massively critical element of data centre architecture. Weak cable implementation is not just unclear it can obstruct airflow, hinder the correct removal of hot air and stop cold air from entering. Cable air damming over time may cause overheating and failure of equipment, resulting in costly downtime. Data centre cabling was usually mounted beneath a high level. Designs have, though, improved in recent years to at least some flexibility for overhead cabling that also helps to lower electricity costs and minimise refrigeration needs. In order to maintain continuity of output and easy usage, well-managed facilities using organised cables procedure. Unstructured point-to-point cables cannot be mounted at all but are also correlated with higher operational

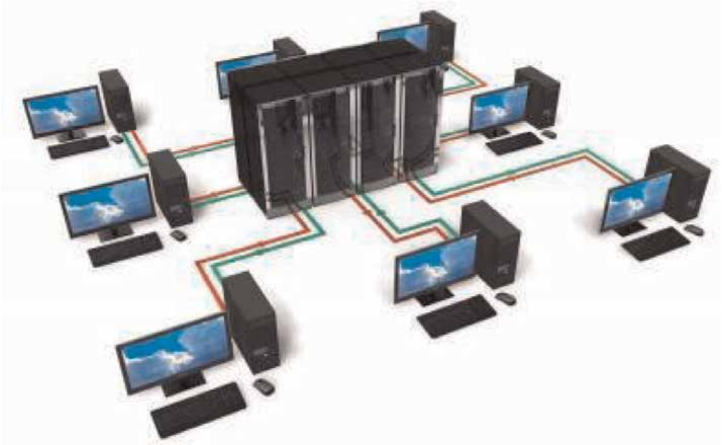


Figure 9.
Network infrastructure 1 [31].

expenses and severe maintenance issues. A successful first move to networking for the food ordering application is the proper cable control [32].

3.4.2 Connectivity

The abundance of ISP networking solutions is one of the key benefits of a carrier-neutral data centre. Basically, a data centre, like any other person, links to the internet: through a different service provider cable. Moreover, in comparison to a traditional house, data centres provide many links with different vendors, enabling the food ordering application a variety of choices. The various networking solutions often provide a lot of flexibility, and it is nearly always easy to reach the external internet. Blended networking solutions often have major protections against DDoS assaults (**Figure 10**) [32].

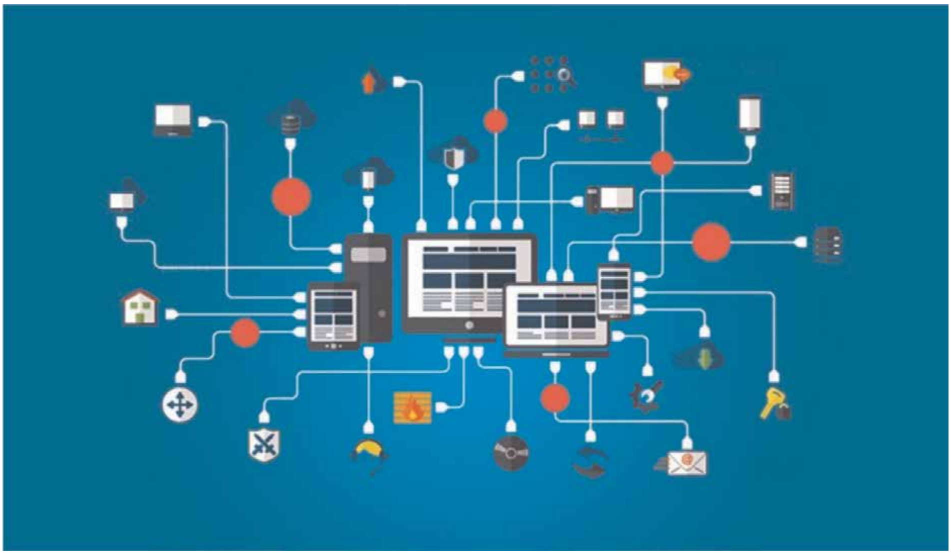


Figure 10.
Network infrastructure 2 [33].

3.4.3 Routers and switches

Cabling from data centres is difficult enough even without routers and improvements to guide network traffic flow through and across the facility will approach nightmarish rates of difficulty. These devices act as unified nodes that enable data to move as easily as possible from one location to another. Properly installed, they can handle vast volumes of traffic and form a vital part of the topology of the data centre without losing efficiency. Incoming public internet data packets first reach the edge-routers of the data centre, which evaluate from where any packet arrives and where it has to go. It transmits the packets to the core routers that form an additional layer at the deployment stage. Such appliances are more aptly defined as switches, as they handle traffic in the data centre networking infrastructure. The grouping stage is named a set of key switches as all traffic is guided inside the datacentres. When it requires data to move through not physically linked computers, it must be transferred through the main switches. Having a wide number of addresses for the core to handle and sacrificing pace becomes necessary for individual servers to contact each other, and the data centre networks prevent the issue by linking server batches to a third layer of switches. Often these classes are called pods which encrypt data packets to allow the core to recognise what which traffic should be guided to rather than handling individual server requests for the food ordering application [32].

3.4.4 Servers

Deployments with high-density servers appear to have higher cabling, cooling and power usage specifications. The food ordering application wants their equipment in racks with convenient access to direct connections and individual cross-connections that provide better efficiency, pace and reduced downtime effect [32].

3.4.5 Direct connection

Perhaps the internet of the data centre is not quick enough to meet the demands of a client. It cannot require the lag or downtime of a cloud service [34] provider link. Data centres may in these situations give them the benefit of directly linking the server with one cross-connection to the servers of the provider. Through the direct cable running between servers, customers may work better when reducing latency and downtime. Although data centre networks are complicated structures which must be carefully maintained to guarantee high-quality efficiency, in any facility the basic framework is focused on exactly the same principles. By improving these networks, data centres will also be an enticing venue for businesses wanting to position their IT systems with a third-party supplier while providing a selection of creative offerings to their clients [32].

4. Cooling system

In most data centres, the server room's temperature is always lower than in other places in the same building. For every single data centre, when the data collected on servers are growing, and a continuous increase in processor performance, it will lead to the servers will produce more heat. When the server's temperature reaches a critical point, it might cause the server not to be able to work properly, and the processor will slow down the performance to avoid overheating. When coming to the worst-case scenario, extremely high temperatures will burn the processor, eventually cause services interruption, and require replacing the hardware for resuming the services. For many reasons showing why a data centre needs a cooling system, and the cooling system is one of the essential components in the data centre. Hence, selecting a cooling system that can operate continuously, and the reliability is the top priority.

The cooling system operates in a variety of ways and has different levels of performance. Generally, there are two main methods for data centre cooling, which are an air-based cooling system and a liquid-based cooling system [35]. In liquid-based cooling system is a reduction of heat from the server by exploiting the properties of liquids [36]. For the air-based cooling system, commonly there are three types, transitional cooling, hot aisle containment, and cold aisle containment. These cooling systems cool down the server room temperature via cold air.

According to the research, the air-based cooling system method has commonly been used for years, and it is simple to compare to the liquid-based cooling system. For a liquid-based cooling system has the risk of leakage across the server rooms, it could damage the components if not appropriately used [37]. Therefore, in this project, we will focus on the air-based cooling system.

4.1 Hot aisle containment system (HACS)

Hot aisle containment system (HACS) composes the hot aisle to collect the hot exhaust air from the server at the back of the racks, and cold supply air is brought in by the equipment at the front. This is similar to a traditional hot aisle/cold aisle arrangement where data centre racks are arranged in alternate rows, cold air intakes facing one way and hot air exhausts facing the other, it keeps hot air in one aisle and cold in the other. While the traditional hot aisle/cold aisle cooling system, which only uses rack placement, worked well in a low-density environment but did not completely isolate the aisles and prevent hot and cold air from mixing [38]. The only method to stop the hot and cold air from mixing up is to form a physical barrier. This is where the containment systems are coming in. Typically, the containment forms a physical barrier from the top of the server racks to the drop ceiling, and it contains the hot aisle, and the exhausted hot air directly back to the cooling units [39]. The cold air will come from the CRAC unit, and the environment outside of the hot aisle becomes a large cold air plenum [40]. This is ensuring the hot and cold air are isolated (**Figure 11**).

Pros of HACS

- Overall the server room is in a cold environment.
- Any leakage of conditioned air will not affect it, as it goes into the cold aisle.
- More effective in cooling.
- Able to use standard fire detection system without having any obstruct.

Cons of HACS

- The cost of the construct is higher.

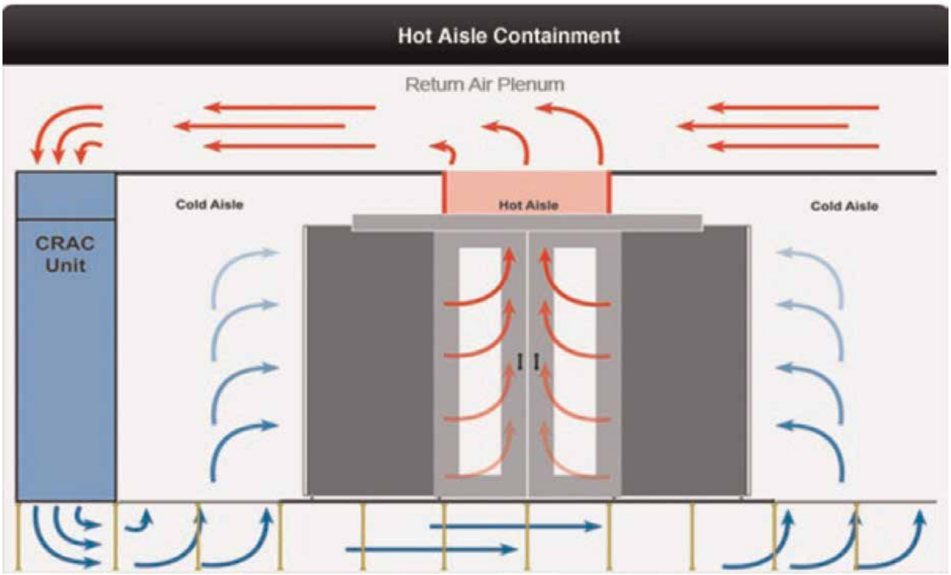


Figure 11.
Hot aisle containment system [35].

- It needs a contained path to return the exhaust air to the cooling unit.
- Hot aisle's temperature easily becomes higher [41].

4.2 Dynamic cooling management and optimisation

Cooling management and optimises system continuously optimise airflow in the data centre equipment, performing improved reliability and availability. The system uses a dense array of temperature sensors to discover precisely where the hot spot is in the data centre. It helps identify potential equipment risks and automatically eliminates up to 95% of hot spots. As the IT load changes, integrated machine learning automatically adapts cooling to varying IT loads to balance the dynamic data centre environment. The system adjusts the cooling need with the lowest possible energy consumption, achieving the immediate cost savings and a suitable amount of cooling in the data centre (**Figure 12**) [42].

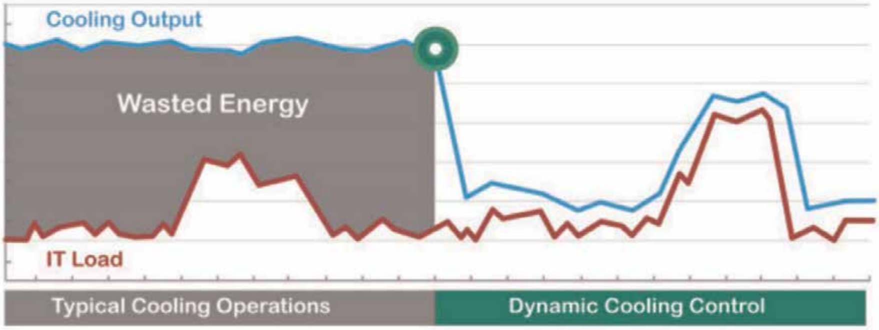


Figure 12.
Dynamic cooling control [42].

4.3 Justification

In order to achieve a tier 4 data centre, all components in the data centre must be fully redundant, including the cooling system. A cooling system design for a tier 4-rated data centre should fulfil the requirements below:

Redundant components—a backup of equipment for a cooling system such as:

- CRAC Units
- Chillers
- Fans

After comparing the air-based cooling system, the hot aisle containment system is the chosen system implement for this project. According to Gavin Banks, HACS is significantly more cost-saving, compare to cold aisle containment system (CASC), it can save 43% in energy cost, which could relate to a 15% reduction in PUE [43]. This is also supported by Schneider Electric, in the report showing HACS provides 40% more energy cost-saving annually compare to CACS [40]. The legacy/traditional cooling system could be more expensive due to inefficiency and uneven cooling, and it may also require more and oversized equipment to accomplish the task. When looking at the bigger picture and all the considerations, other IT equipment in the same room that would need cooling as well, cold aisle containment system might make the server room area extremely hot. It will become a challenge for those who require to be inside the server room for maintenance or servicing, as well as for other IT equipment work under high temperatures. Therefore, hot aisle containment would be the appropriate choice (**Figure 13**).

4.4 Space allocation

A proposed layout is as follows (see **Figures 14** and **15**).

4.5 Physical security

A breach of physical security may cause unimaginable damage to a data centre. Given the growing need to protect valuable information, any loss of data or even the

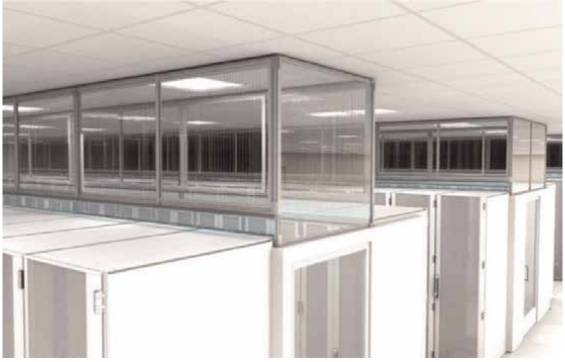


Figure 13.
Hot aisle containment [39].

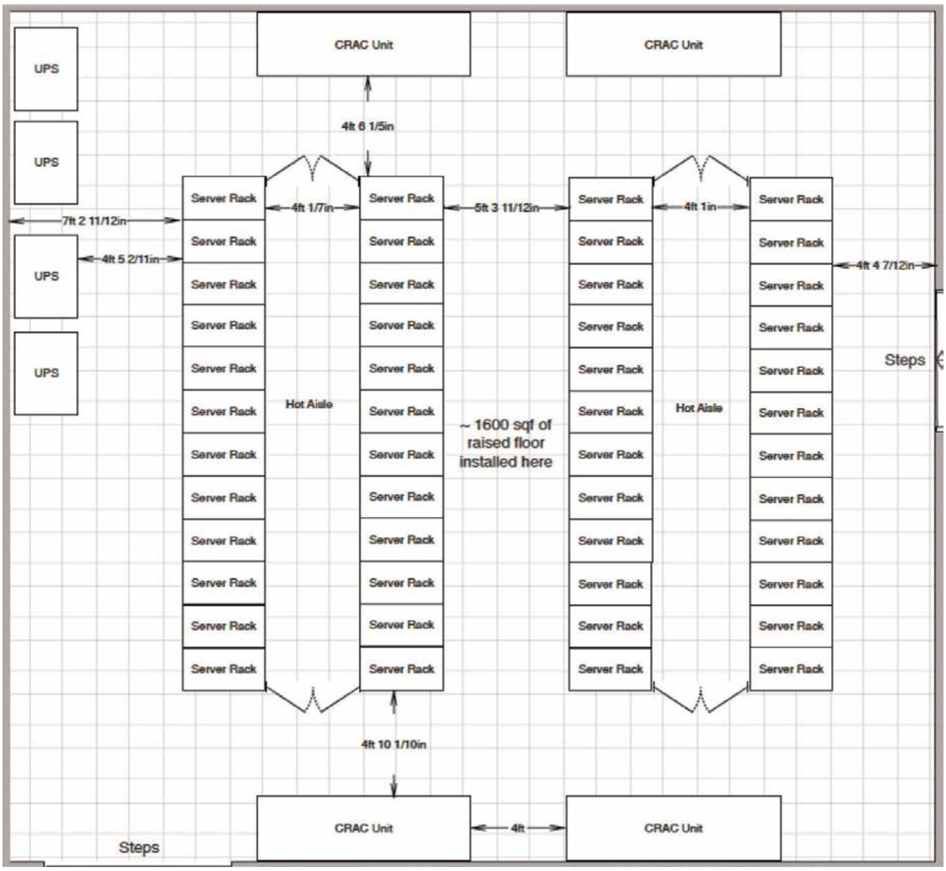


Figure 14.
Proposed layout [44].

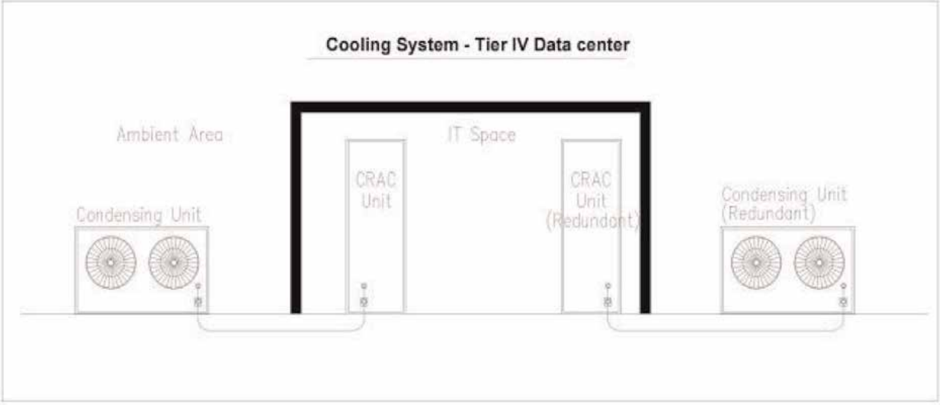


Figure 15.
Tier 4 N + N cooling system [45].

incapability to comply with mandatory regulatory requirements may result in obloquy, loss of customers, fines and loss of revenue. Interoperability is a critical building block for the physical security of a data centre. The entire ecosystem of manufacturers and integrators serving the data centre physical security market needs to ensure that the products work together to provide a scalable, layered physical security solution [46].

The prime purpose of implementing physical security is to protect the information, devices and IT infrastructure of the data centre from any threat that could disrupt the operation of the data centre. It could be caused by any illegal activity, such as theft, leakage of data or damage by any physical involvement in the data centre. Building a layered approach to data centre security helps to customise the solution to the needs of a data centre. The organisation needs to determine the right layered approach, and understand the current system, the working environment and future needs (**Figure 16**).

A practical, layered approach requires all systems to function coherently. Generally, the security architecture consists of multiple layers of physical security that need to be considered to protect the data centre as a whole and to comply with the data centre protection guidelines (**Figure 17**).

4.5.1 Layer 1: perimeter defence

The first layer of physical security is perimeter defences, a physical boundary or fence at the property edge to deter external threats, which is controlling and restricts the access to the data centre property. There are three D's to describe the purpose of perimeter security, Deter, Detect, and Delay [49]. Usually, there are only two doors that are allowed to enter the data centre, the front door and the loading bay. The perimeter fence detection system can integrate with trespassing alarms, high-definition CCTV system, limited access control points, and motion-activated security lighting.

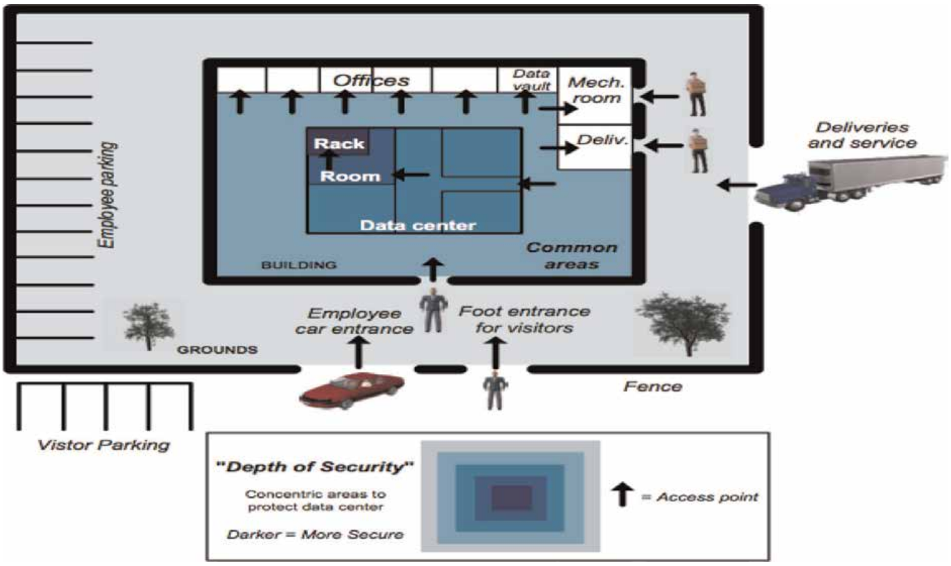


Figure 16.
Security map showing the depth of security [47].

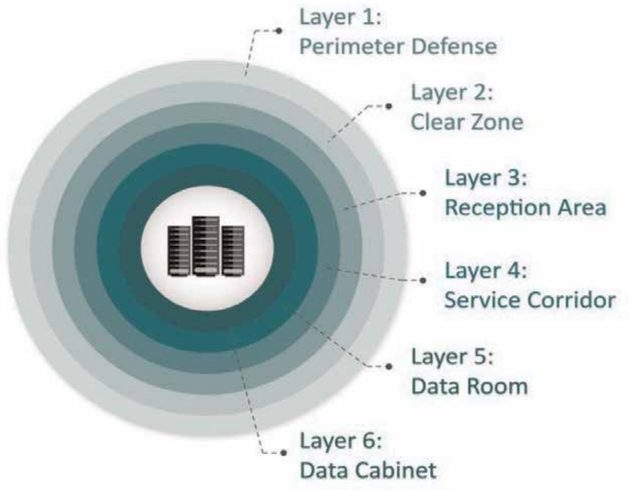


Figure 17.
Layers of physical security in data Centre [48].

4.5.2 Layer 2: clear zone

The second layer called as Clear Zone creates a buffer zone between the perimeter, and the data centre to have better detect physical invasion [48]. The clear zone is also a large area containing critical infrastructures such as fuel containment, generators, and main power supplies [46]. This zone needs security measures that provide a total awareness of the situation.

4.5.3 Layer 3: facility entrance and reception

Control visitor's access to the data centre and validate authorised access. All the employees and visitors before entering the data centre must check-in or register at the front desk, visitors, must obtain a temporary pass in order to access the secured areas.

4.5.4 Layer 4: services Corridor (Escorted areas and Grey space)

Validate the rights of authorised persons to access specific areas within the building. The corridors, grey spaces and escorted areas that head to the data centre floor are often where the proper security measures are overlooked [46]. This could lead to unauthorised access to critical mechanical and electrical infrastructure.

4.5.5 Layer 5: data centre room

In this layer, it is further restricting access through various forms of authentication, and monitoring all authorised access. Implement high-security electronics to prevent general staff or trespassers from accessing sensitive areas. To prevent unauthorised persons from entering white space, access control such as dual-factor biometrics is essential for the control of authorised access to the data centre.

4.5.6 Layer 6: date centre cabinet

Establish the protection of sensitive electronics (servers) that contain crucial data. The security measures to accomplish include cabinet locking, audit trails and an intelligent infrastructure strategy. This layer is especially essential and effective in reducing the critical and frequently forgotten the insider threat.

Most data centres did an excellent job of accomplishing the first few layers. Still, the absence of reliable control of the cabinet may result in costly data breaches caused by a malicious or disgruntled employee, or maybe even unknowing and unintentional access to data.

5. Synthesis

Power: The data centre for the food ordering application will be using the grid electricity as the main power source to power up all the equipment and components in the data centre. In case of a power break or power down all the equipment's will be powered with the UPS backup battery for a short time and by that time the diesel generator will replace the power source with it. For every data centre uptime for its servers are highly crucial since the clients should have uninterrupted services. And 99.9% uptime will ensure uninterrupted services for all the customers and keep the data secure and ongoing.

5.1 Server racks and computing resources

The server racks will consist of all the necessary components for the servers to compute the desired function. Which will process placing an order, processing the payment and creating data for the customer. Which should be done within no time and should keep things fast are reliable. The servers are responsible for doing all the backend processes of the application and should be running all the time without any issues. All of the components should be working properly. And should be connected with each other to compute the tasks.

5.2 Storage infrastructure

Every customer detail and data should be kept in a storage device. The storage system consists of many storage components which will be connected to the servers all the time and store the necessary data in the storage. They will compute the data to the server interaction with it. Hard disk drive, tape drive and other forms of internal and external storage devices make process and stores the computed data. There will be a backup for all the data in case of storage devices get interrupted and to keep it working depending on necessary times. The storage utility software keeps monitoring [50] the process for uninterrupted processes.

6. Evaluation

6.1 PUE and efficiency

Data centres use significant amounts of power to operate. Majority of the power is consumed by the cooling systems. A successful data centre must be efficient. One of

Components	Qty	Power (W)
IT		
Compute Rack	8	8 × 5250 W = 42,000 W
Storage Rack	2	2 × 5250 W = 10,500 W
Switch (Fibre Channel and Ethernet)	10	10 × 300 W = 300 W
Router	4	4 × 500 W = 2000 W
Computer	2	2 × 300 W = 600 W
Printer	1	1 × 250 W = 250 W
CCTV Camera	14	14 × 10 W = 140 W
Total		55,790 W
Non-IT		
Cooling	1	1 × 40000 = 40,000 W
Lighting	20	20 × 50W = 100 W
Smoke Detectors/Alarms	14	14 × 2W = 28 W
Misc.	1	1 × 500W = 500 W
Total		40,628 W
Total Power for Facility		55,790 + 40,628 = 96,418 W

PUE = Total Power Consumption/IT Energy Needs.

Table 1.
Estimated power usage for different components and equipment of the data centre based on enterprise IT equipment.

the metrics for calculating the efficiency of the data centre is Power Usage Effectiveness (PUE). PUE is calculated by using the total amount of power consumed by the energy used by the IT equipment [51, 52].

In order to calculate the PUE, values for power consumption and IT energy needs are to be determined. The **Table 1** shows the estimated power usage for different components and equipment of the data centre based on enterprise IT equipment.

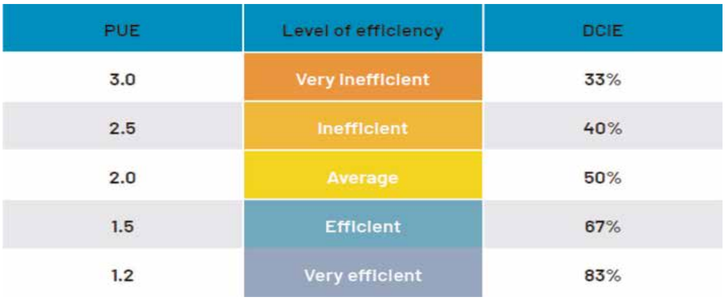
Based on the estimates above, the PUE for the data centre is calculated by Eq. (1),

$$\mathbf{PUE} = 96{,}418/55{,}790 = \mathbf{1.73}. \tag{1}$$

The PUE value lies between efficient and average according to **Figure 18**. One of the main reasons why the efficiency is slightly below efficiency is that cooling system requires more energy due to the geographical location of Malaysia. In other countries such as Iceland, the cooling system will not need that much power because they are also closer to the poles and they experience the winter season which Malaysia does not. However, the efficiency is still optimal for a data centre in this region.

6.2 Expandability

A data centre is expected to meet future business needs and expand accordingly. The food company already has 5 million users, in the next few years, they estimate their user base will increase given the current popularity. Therefore, the data centre design must be able to take this into consideration and allow future expansions.



PUE	Level of efficiency	DCIE
3.0	Very Inefficient	33%
2.5	Inefficient	40%
2.0	Average	50%
1.5	Efficient	67%
1.2	Very efficient	83%

Figure 18.
Level of efficiency relative to the PUE value [37].

According to [53], many data centre expansions result in failure. As for expandability in this data centre, it will initially occupy 10 racks which are about 25% of the floor space available to avoid the mistake of oversizing and wasting resources. This allows the facility to not be overcrowded when demands rise. Starting with few racks lowers the cost to build (CapEx). In addition to this, the usage of rack enclosures/cabinets helps with cooling. Therefore, in the future cooling systems will not be overburdened with additional racks in the facility since rack enclosures have better airflow and cooling. The data centre is designed with a modular approach. Designs that are modular and flexible are the key to long-term success [53]. For example, increasing the storage capacity is trivial since the data centre storage infrastructure is based on Storage Area Network (SAN) which offers great scalability and expandability compared to other architectures. Finally, through proper planning using the total cost of ownership (TCO) approach and flexibility of the facility, the data centre can meet the requirements of recent market demand.

7. Conclusion

We may sum up by noting that because the IT industry is continually expanding, there will always be a need for new and improved solutions. There is no doubt that the solution and tools selected for this work will remain the same in the future. The intended data centre is meticulously thought out, from security to smart execution. Considerations for the design of a data centre in terms of power efficiency, cooling systems and protection should include scalability, power effectiveness, CO2 reduction, system resilience, sustainability, the use of machine learning, and other cutting-edge technology. Additionally, by renting space in the data centre and selecting their own equipment, clients using the co-location system can locate their data. Finally, through proper planning using the total cost of ownership approach and flexibility of the facility, the data centre can meet the requirements of today and tomorrow.

Conflict of interest

The authors declare no conflict of interest.

Author details

Yaseein Soubhi Hussein[1]*, Maen Alrashd[2], Ahmed Saeed Alabed[1] and Amjed Zraiqat[3]

1 Computer Science and Information Systems Department, Ahmed Bin Mohammed Military College, Qatar

2 Faculty of Science and Information Technology, Jadara University, Irbid, Jordan

3 Faculty of Science and Information Technology, Al-Zaytoonah University of Jordan, Amman, Jordan

*Address all correspondence to: dr.yaseein@abmmc.edu.qa

IntechOpen

References

[1] Datacenters.com. Everything You Need to Know About Data Center Power. 2020. [Online]. Available from: https://www.datacenters.com/news/everything-you-need-to-know-about-data-center-power

[2] Kutsmeda K. Data Center Power Strategies. 2013. [Online]. Available from: https://www.csemag.com/articles/data-center-power-strategies/

[3] Shaw R, Howley E, Barrett E. Applying reinforcement learning towards automating energy efficient virtual machine consolidation in cloud data centers. Information Systems. 2022;**107**:1-21

[4] Lorbel. What is power distribution unit & automatic transfer switch?. 2018. [Online]. Available from: https://www.lorbel.com/what-is-power-distribution-unit-automatic-transfer-switch

[5] Osemco. Closed Transition Transfer Switch, [Online]. 2016. Available from: http://osemco.co.kr/en/pro/closed-transition-transfer-switch.html

[6] Evanuik S. UPS in Critical Data Centers. 2019. [Online]. Available from: https://www.allaboutcircuits.com/technical-articles/uninterruptible-power-supply-systems-in-critical-data-centers/#:~:text=The%20uninterruptible%20power%20supply%20(UPS,disturbances%20and%20power%20quality%20issues

[7] Bergum M. Line Interactive vs. Double Conversion UPS—Which One's Best?. 2019. [Online]. Available from: https://www.qpsolutions.net/2019/11/line-interactive-vs-double-conversion-ups-which-ones-best/

[8] Woodstock Power. Backup generators for data centers. 2019. [Online]. Available from: https://woodstockpower.com/blog/backup-generators-for-data-centers/#:~:text=Backup%20generators%20for%20data%20centers%20provide%20power%20when%20the%20main,high%20risk%20of%20operational%20loss

[9] Vertiv. What is a Rack PDU?, 2020. [Online]. Available from: https://www.vertiv.com/en-emea/about/news-and-insights/articles/educational-articles/what-is-a-rack-pdu/#:~:text=Rack%20Power%20Distribution%20Units%20(rPDUs,equipment%20within%20the%20data%20center.&text=The%20rPDU%20then%20distributes%20power,each%20ind

[10] Data Center Power Chain: How It Works. Somerset, New Jersey, NJ, USA: Raritan; 2018

[11] Rundquist R. What Type of Rack PDU is Right for Your Data Center?. 2019. [Online]. Available from: https://www.vertiv.com/en-emea/about/news-and-insights/articles/blog-posts/what-type-of-rack-pdu-is-right-for-your-data-center/

[12] Bord M. On Switched Rack PDUs and Decreasing Energy Consumption in the Data Center. 2014. [Online]. Available from: https://www.raritan.com/blog/detail/on-switched-rack-pdus-and-decreasing-energy-consumption-in-the-data-center

[13] Mitsubishi Electric. Double Online Conversion UPS Technology [Online]. 2020. Available from: https://www.mitsubishicritical.com/technologies/double-conversion-vs-line-interactive/#:~:text=Online%20double%20conversion%20UPS%20systems,(PUE)%20increase%20their%20ROI

[14] DPS. How a rack switched PDU protects your remote sites, [Online].

2022. Available from: https://www.dpstele.com/power-distribution-unit/switched-rack-pdu-definition.php#:~:text=The%20Strong%20ROI%20of%20a%20Switched%20Rack%20Power%20Distribution%20Unit&text=With%20switching%20functionality%2C%20a%20remote,or%20reboot%20a%20remote%20device

[15] Office of Energy Efficiency & Renewable Energy. Energy 101: Energy Efficient Data Centers, 2020. [Online]. Available from: https://www.energy.gov/eere/videos/energy-101-energy-efficient-data-centers#:~:text=Data%20centers%20can%20become%20more,and%20help%20protect%20the%20nation

[16] Schroll RC. Fire Detection and Alarm Systems: A Brief Guide. 2007. [Online]. Available from: https://ohsonline.com/Articles/2007/12/Fire-Detection-and-Alarm-Systems-A-Brief-Guide.aspx

[17] Walker T. Three Level of Data Center Fire Protection. 2019. [Online]. Available from: https://www.firetrace.com/fire-protection-blog/three-levels-of-data-center-fire-protection

[18] IFSEC GLOBAL. Smoke Detectors Explained. 2019. [Online]. Available from: https://www.ifsecglobal.com/smoke-detectors/

[19] Menke J. Air Sampling Smoke Detection. 2016. [Online]. Available from: https://blog.a1ssi.com/air-sampling-smoke-detection/

[20] Kaiser L. What is an Air Sampling Smoke Detection System?. 2015. [Online]. Available from: https://www.orrprotection.com/mcfp/blog/air-sampling-smoke-detection-system#:~:text=Air%20sampling%20detectors%20are%20chosen,at%20air%20handling%20return%20grilles.&text=Buildings%20where%20other%20smoke%20detectors%20have%20failed

[21] Honewell Gent. Products Aspirating Smoke Detection. [Online]. 2019. Available from: https://www.gent.co.uk/products/aspirating-technology/#:~:text=Honeywell%20Gent's%20range%20of%20Aspirating,and%20before%20intense%20smoke%20develops

[22] Crimmins D. What is a Fire Alarm System?. 2019. [Online]. Available from: https://realpars.com/fire-alarm-system/

[23] London Fire Brigade. Fire Alarms. Learn about Different Kinds of Fire Alarm Systems, and Get a Better Understanding of What You Might Need to Instal in Your Property. 2020. [Online]. Available from: https://www.london-fire.gov.uk/safety/property-management/fire-alarms/

[24] Kennedy K. Wired Vs. Wireless Fire Alarms—What's the Best Choice for Your Business?. 2018. [Online]. Available from: https://www.sshfireandsecurity.co.uk/news/wired-vs-wireless-fire-alarms-whats-the-best-choice-for-your-business#:~:text=Addressable%20systems%20are%20therefore%20more,control%20panel%20using%20radio%20waves

[25] BOSCH. 2020. [Online]. Available from: https://www.boschsecurity.com/gb/en/solutions/fire-alarm-systems/

[26] Discount FireSupplies. Wireless Fire Alarms. [Online]. 2020. Available from: https://www.discountfiresupplies.co.uk/category/22/Wireless-Fire-Alarms#:~:text=Wireless%20fire%20alarm%20systems%20are,existing%20wired%20fire%20alarm%20systems

[27] FireSystems INC. Why your Business needs a Wireless Fire Alarm

System. 2019. [Online]. Available from: https://firesystems.net/2019/03/16/why-your-business-needs-a-wireless-fire-alarm-system/

[28] PROEN Internet. PROEN Internet, 2019. [Online]. Available from: http://www.siamidc.com/firesuppression.php

[29] Walker T. Firetrace.com, 2020. [Online]. Available from: https://www.firetrace.com/fire-protection-blog/three-levels-of-data-center-fire-protection

[30] Ghaffari E, Rahmani AM, SaberiKamarposhti M, Sahafi A. An optimal path-finding algorithm in smart cities by considering traffic congestion and air pollution. IEEE Access. 2022; **2022**:55126-55135

[31] Network Infrastructure. 2019. [Online]. Available from: Trustednetworksolutions.com

[32] Alan Seal. Vxchnge.com, 2019. [Online]. Available from: https://www.vxchnge.com/blog/data-center-networking-101-everything-to-know

[33] 1Connect SOHO. 1 Connect SOHO, 2020. [Online]. Available from: https://1connectsoho.com/upgrade-network-infrastructure-asap/

[34] Sadri AA, Rahmani AM, Saberikamarposhti M, Hosseinzadeh M. Data reduction in Fog computing and internet of things: A systematic literature survey. Internet of Things. 2022;**2012**:1-31

[35] Isberto M. The Latest Innovations in Data Center Cooling Technology, 2018

[36] Rouse M. What is Liquid Cooling?—Definition from WhatIs.com. [Online]. Available from: https://whatis.techtarget.com/definition/liquid-cooling

[37] Mezzanotte M. How to Calculate the PUE of a Datacenter. 2019. [Online]. Available from: https://submer.com/blog/how-to-calculate-the-pue-of-a-datacenter/

[38] B. N. Technologies. 2018. [Online]. Available from: http://www.bluewavenetwork.net/data-centers.html

[39] Shield C. Hot Aisle Containment HAC Systems For Data Centers, 2016

[40] Avelar k. Brown. Impact of Hot and Cold Aisle Containment on Data Center Temperature and Efficiency, 2011

[41] Ahdoot A. Is cold or hot aisle containment better for your data center? 2014. [Online]. Available from: https://www.colocationamerica.com/blog/hot-vs-cold-aisle-containment

[42] Schneider Elecrtic. StruxureWare data center operation, 2016. [Online]. Available from: www.apc.com/struxureware

[43] Banks G. Advantages of Hot Aisle vs. Cold Aisle Containment, 2019. [Online]. Available from: https://www.sourceups.co.uk/hot-aisle-vs-cold-aisle-containment/

[44] ColocationPLUS. Take the Red or the Blue? How We Chose Hot Aisle Containment for Our Newest Data Center—ColocationPLUS. [Online]. 2019. Available from: https://colocationplus.com/2018/10/16/take-red-blue-chose-hot-aisle-containment-newest-data-center/

[45] Moumiadis T. Tier 4 data center cooling system design—My engineering notes. [Online]. 2019. Available from: http://moumiadis.blogspot.com/2019/03/tier-4-data-center-cooling-system-design.html

[46] Anson S. Data Center Security Best Practices | Security Info Watch. 2017. [Online]. Available from: https://www.securityinfowatch.com/perimeter-security/physical-hardening/article/12336001/datacenter-security-best-practices

[47] Niles S. 2011. [Online]. Available from: https://securitytoday.com/-/media/1837DD6DB5F441D69DC863B6D7EEF94A.pdf

[48] ICD. How to Secure Your Data Center—ICD Security Solutions, ICD Security Solution. [Online]. Available from: https://www.icdsecurity.com/2019/08/28/how-to-secure-your-data-center/

[49] Anixter. The Four Layers of Data Center Physical Security for a Comprehensive and Integrated Approach. 2012

[50] Fathi A, Hussein YS, Sabri NA. Leverage networking tasks using network programmability. TEST Engineering & Management. 2020;**83**: 905-910

[51] Felter B. What is Power Usage Effectiveness and How it Impacts Your Costs. 2020. [Online]. Available from: https://www.vxchnge.com/blog/power-usage-effectiveness

[52] Manaserh YM, Tradat MI, Bani-Hani D, Alfallah A, Sammakia BG, Nemati K, et al. Machine learning assisted development of IT equipment compact models for data centers energy planning. Applied Energy. 2022;**2022**:1-17

[53] Hagan MM, Lusky J, Hoang T, Walsh S. 2016. [Online]. Available from: https://download.schneider-electric.com/files?p_File_Name=VAVR-8K4U25_R1_EN.pdf

Section 2

Applications and Tools

Chapter 3

Ontology-Based Solution for Handling Safety and Cybersecurity Interdependency in Safety-Critical Systems

Dionysia Varvarigou, David Espes and Giacomo Bersano

Abstract

In case, safety-critical systems face an anomaly (either intentional or not), safety and cybersecurity impact humans and environment. Thus, they affect each other and so they are considered as interdependent. An ontology-based solution for safety is needed to handle this interdependency. We propose a new safety ontology for Network Function Virtualization (NFV) framework which is able to cover reliability, availability, maintainability, and integrity-related breakdown types, since they interact and influence safety according to ENISA. Our ontology allows us to have a uniformized representation of the potential anomalies that a system and its elements can face. Based on this representation, a decision-making process takes place to avoid potential conflicts between safety and cybersecurity in order to best handle their interdependency. The results of our implementation show that our ontology handles the safety and cybersecurity interdependency and has little impact on decision-making time, which makes it an effective methodology for NFV framework.

Keywords: safety ontology, NFV safety architecture, safety and cybersecurity interdependency, Network Function Virtualization (NFV)

1. Introduction

In safety-critical systems, safety is the most significant property to be considered. This is because the main focus for these systems is to prevent harm on humans and environment. However, safety is able to interact with other properties as well. According to the ENISA standard [1], safety is a subset of the reliability, maintainability, availability, and integrity properties. In this way, it is understood that safety has the ability to interact with these aforementioned properties while considering their impact to humans and environment. Furthermore nowadays, NFV applications are expanded as they are used to various types of systems. Thus, NFV can be applied in safety-critical systems. In this case, safety is an important property for NFV. In [2], an NFV application is used in a safety-critical use case which proves the importance of safety in these systems. For example, NFV handles services for an autonomous vehicle. In case, a reliability anomaly happens in one of the NFV services and the vehicle becomes uncontrollable, it can have an impact on the people that it carries, the people in the surrounding, and the surrounding environment itself.

However, as seen in ENISA standard, the properties that interact with safety are shared with some of the properties of cybersecurity. This makes understood that safety is also able to interact with cybersecurity. Thus, the functionalities of safety are able to influence and violate the ones of cybersecurity. Likewise, the functionalities of cybersecurity can affect the ones of safety. As an outcome, it is possible to consider safety and cybersecurity as interdependent. As an example, in order for cybersecurity needs to mitigate an anomaly, it asks for a re-launch of a Virtualized Network Function (VNF). This is issued to the NFV Orchestrator (NFVO) module. This is because this module is the responsible one for implementing all the issued orders. At that moment, safety understands that this action goes against its safety measures and blocks the NFVO from issuing this specific re-launch.

In order to prevent any safety and cybersecurity violations, it is needed to be able to differentiate the safety anomalies from the cybersecurity ones. An ontology-based solution is a good way to automate this process. Thus, it is possible to find ontology-based solutions for each one of the safety-related properties independently. However, in the literature, there are no ontology-based solutions for safety considering all the properties related to it as a whole. Moreover, there are no ontology-based solutions that provide a safety and cybersecurity interdependency. This has the effect of limiting the decision-making process that is used for distinguishing the anomalies created in a system. Furthermore, this prevents from taking into consideration the interdependency of safety and cybersecurity.

Thus, it is understood that in order to ensure safety in a NFV framework, there are specific challenges to be addressed. These challenges deal with: (i) the detection and mitigation of a variety of safety anomalies in a more comprehensive way, and (ii) the management of safety and cybersecurity interdependency. In order to handle safety in a NFV framework, an orchestrator is needed which is able to detect reliability, availability, integrity, and maintainability-related anomalies with respect to safety. Ontology is a good option for addressing this issue, since ontology is an explicit specification of a conceptualization where the knowledge of a domain is represented in a declarative formalism [3]. This makes it possible to represent the different types of anomalies in relation to safety. According to this uniformized representation, the reasoner (piece of software) is able to infer logical consequences. These consequences make it possible to understand whether a safety-related anomaly is also a cybersecurity-related one.

To this end, our solution proposes (i) a new ontology for ensuring safety in NFV framework and (ii) specific rules to be used by the reasoner. Our proposed ontology is used by an orchestrator that handles safety in a NFV framework. This ontology includes (i) the description of safety and the properties related to it (i.e. reliability, availability, maintainability, and integrity) as classes, (ii) the concerned elements for each property as subclasses, and (iii) the breakdown types for the potential adversities as object properties. Our proposed rules allow us to automate the decision-making process. This is because the reasoner needs the rules to make a decision. According to this decision, a NFV safety orchestrator is able to modify the plan of mitigation. With this modification, it is possible to avoid potential safety and cybersecurity conflicts.

The remaining of this paper is organized as follows. Section 2 reviews the relevant works of ontologies. Section 3 introduces our proposed ontology. Section 4 provides the rules for supporting the decision-making process. Section 5 presents the evaluation of the feasibility of our proposed ontology. Section 6 provides the results and their analysis. Finally, the last section concludes the study, and it discusses possible future work.

2. Related work

In general, ontologies are used for system modeling, since they are capable of describing a whole system with its components and subsystems. This is because an ontology is expressed as the study of what exists in a certain context [3].

2.1 Ontologies for safety-related properties

As follows, it is possible to provide the related work with respect to ontologies for all the safety-related properties but also the integration of safety and cybersecurity. These ontologies are provided in general.

2.1.1 Safety ontologies

In relation to safety, ontologies are commonly used for obtaining safety risk knowledge and handling safety management. For safety risk knowledge, it is possible to develop an ontological method which organizes this knowledge into seven unified classes (i.e. project, construction activity, risk factor, risk, risk grade, risk consequence, and risk prevention measure) [4]. For handling the risk management, an ontology with a case-based reasoning is used as a decision-making approach for safety risk management [5]. Moreover, safety ontologies are able to represent specifically extracted information from databases. In this way, ontologies can assist for identifying additional capabilities of these information [6].

However, ontologies can be integrated with other technologies, algorithms, or methodologies in order to enhance their capabilities. For instance, ontologies can be integrated with computer vision algorithms to develop knowledge graphs that can automatically and accurately recognize hazards even when they are subjected to change [7]. Another example is when ontologies can be combined with wireless networks to identify potential hazards [8].

2.1.2 Reliability ontologies

Reliability with respect to ontologies is expressed as a way to make ontologies reliable, or to use ontologies for increasing reliability in various systems. Agile methodology uses agile principles and practices for ontology development. In this way, it is possible to utilize software engineering to build reliable ontologies [9]. Moreover, ontology alignment is a way to create reliable ontologies. In [10], machine learning techniques are used to automatically align ontologies to make them more reliable.

However, ontologies are able to be used in various methodologies in order to provide a variety of different types of reliability. In general, an ontology-based text mining methodology is able to maximize system reliability, since it is able to extract knowledge from databases [11]. There are many technologies and methods in order to use semantic web and ontologies for providing reliable services. This is because the use of semantic technologies in the modeling of a multi-agent system are very effective in increasing coordination and interoperability, as seen in [12]. Furthermore, ontologies are able to assist into making the numerical simulation techniques more reliable. This can happen with ontology-based text and data mining techniques, as seen in [13].

2.1.3 Availability ontologies

Ontologies for ensuring availability are not widely researched in the literature, up to our knowledge. However, in [14], ontologies are used to provide and ensure heterogeneous knowledge for a specific concept. By combining these ontologies with optimization algorithms, it is possible to provide high data availability.

To sum up, availability is closely linked to reliability and maintainability. Once a system is reliable and maintainable, then it is possible to satisfy availability [15].

2.1.4 Maintainability ontologies

Maintainability is an attribute that is included in dependability. In order to be able to understand all attributes of dependability but also to compare them, it is possible to use a dependability rating ontology [16]. Thus, it is possible to obtain knowledge about the attribute of maintainability but also in relation to the other attributes. Moreover, ontologies can be created by extracting them from other ontologies or by creating them from scratch. The approach to develop an ontology is able to affect the maintainability. Thus, the evaluation of the ontology development is very important. In [17], the authors propose a methodology for evaluating ontology development from scratch.

Furthermore, it is important to be able to create maintainable ontologies. For achieving this, a methodology is proposed in [18] which is able to construct ontologies using a template-based approach for ontology modeling and instantiation. However, ontologies can be also used to enhance maintainability in a system. In [19], an ontology model is proposed to facilitate maintenance strategies selection and assessment. And in [20], ontologies are used for data accessing in order to enhance system maintainability.

2.1.5 Integrity ontologies

Ontologies can be used for ensuring integrity in a system. This can happen with a framework that is able to leverage an ontology to provide representation of semantically enriched data, as seen in [21]. It is also important to be able to evaluate the ontologies with regard to integrity. In [22], an ontology-based evaluation system is proposed which is a new ontology framework of leverage knowledge modeling. This creates an easy-to-use tool for quantitative identification for integrity by combining ontology and semantic web rule language rules.

However, ontologies need some constraints in their analysis in order to be able to focus on certain attributes. One of the ways that ontology accesses data is by querying via query translation. However, constraints in general in this way of accessing data is not represented. For this reason in [23], a framework for querying data that exploits information with regard to integrity constraints is proposed for ontology-based data access. It is also possible to extend the ontology-based data access into including integrity constraints, as seen in [24].

2.1.6 Confidentiality ontologies

Specifically, ontologies dedicated to confidentiality are not widely researched in the literature, up to our knowledge. However, confidentiality can be found in the ontologies that cover all attributes of the cybersecurity approach of Confidentiality, Integrity, and Availability (CIA). In [25], an ontology is developed that targets a requirement-based threat analysis. These requirements refer to the attributes of CIA, where confidentiality is included.

2.1.7 Safety and cybersecurity ontologies

Safety and cybersecurity are two different concepts, and so their ontologies are composed of different elements and objects. In [26], it is attempted to link safety and cybersecurity objectives in an ontology in order to gain better theoretical understanding.

In order to build ontologies, it is possible to extract them from already existing ones and then expand them. In this way, safety ontologies can be expanded to include also cybersecurity. In [27], an ontology that already represents safety is expanded to consider also cybersecurity for the early stages of a system life cycle. Like this, it is able to gather and rank operational needs, assess the feasibility of the desired solution, and pinpoint any technological gaps. Moreover, in [28], a functional safety ontology is improved to consider attack scenarios. In this way, an ontology-based model for functional safety and cybersecurity verification and validation is proposed.

Finally, [29] attempts to integrate safety and cybersecurity in an ontology. This is different from the previous because the previous expand an already existing ontology to consider also cybersecurity, and they consider the early stages of a system's life cycle or the verification and validation process. While, this safety and cybersecurity ontology that is based on formal methods is able to represent the reaction of the system in different kind of scenarios.

2.2 Cyber-Physical Systems

With the use of ontologies, it is possible to understand the relationships between components whether they are cyber or physical ones. In [30], an ontology framework is able to capture the relationships between cyber and physical systems. Ontologies have a wide range of usage, since they can be used as analysis tool and a way to build knowledge hubs. For the analysis tool usage, the Technology Function Matrix is developed based on ontologies [31]. In order to build a knowledge hub, the authors in [32] use an ontology-based structure.

2.2.1 Safety properties

In relation to safety and in order to develop an ontology which considers all the properties that it interacts with as a whole, it is needed to understand how each relevant work provides partial coverage of the safety properties. Starting from maintainability, OntoProg is an ontology-based solution which is used for correct decision-making and assisting in the implementation of the Prognostics Health Management, for mechanical machines [33]. Furthermore, adding also the availability property to maintainability, an ontological structure is provided for availability as a criticality analysis which determines the maintenance strategy [34].

In [35], the three properties of reliability, availability, and maintainability, are provided. However, each one of these properties are found in a different super-concept of the solution, which means that they are not associated. Finally, the reliability and availability properties are provided through an ontological solution for detecting and preventing the failures of the system components of Cyber Physical Systems (CPS) [36]. An ontology is used with all the CPS failures described in order to assist a multi-agent architecture to detect and identify the potential failures. And in [37], an ontology is built by transforming the results of the Failure Modes, Effects, and Criticality Analysis model into a class diagram. This ontology is utilized for detecting and preventing failures. As seen from above, the only paper that is the closest to the global image of safety is the paper that includes availability, maintainability, and reliability [35]. This is because it is the only solution that includes three of the properties that interact with safety.

2.2.2 Confidentiality property

Up to our knowledge, confidentiality ontologies for CPS are not widely researched in the literature. However, since confidentiality is a subproperty of dependability according to ENISA, it is possible to find ontologies that consider confidentiality for CPS in ontologies that concern all attributes of dependability. In [36], an ontology that concerns all attributes of dependability is used to consider various failures.

Additionally, confidentiality is also a subproperty of trustworthiness according to ENISA. Thus, it is possible to find ontologies that consider all attributes of trustworthiness. In [38], SIMON is an ontology framework that is able to ensure trustworthiness and by extension all of its attributes.

2.2.3 Safety and cybersecurity interdependency

In order to build an ontology that handles the safety and cybersecurity interdependency, it is needed to see if there are any research papers in the literature that cover this topic. However, in the literature, there are no papers for safety and cybersecurity interdependency in relation to CPS. In the literature, most of the papers for trustworthiness in CPS use the NIST CPS [39] standard, and none of them is using the ENISA one. In NIST CPS, safety, security, and reliability are subgroups of trustworthiness, while cybersecurity with the CIA approach are subgroups of security. For example, in [40], a framework is provided for reasoning about NIST CPS trustworthiness in CPS, which combines ontology-based reasoning and answer set programming. And in [41], an ontological design and verification framework is presented, which captures the relationships between cyber and physical components in CPS. Once again, NIST CPS trustworthiness is considered.

Furthermore, there is also STRAM, which is one more framework for trustworthiness [42]. According to STRAM, security and trust are its subgroups. Safety and reliability are subgroups of trust, while cybersecurity is a subgroup of security. Both NIST CPS and STRAM consider all of our properties separately and do not associate them. Moreover, in line with NIST and STRAM, safety and cybersecurity share no common properties. This makes us understand that by using NIST CPS or STRAM, there is no way to associate safety and cybersecurity in an interdependent way. However, ENISA gives us an image of the properties that interact with safety, as well as the properties that interact with cybersecurity. Moreover, ENISA also shows the two shared properties between safety and cybersecurity, according to which it is possible to build an architecture that provides safety and cybersecurity as interdependent.

Up to our knowledge, it is possible to distinctively find ontologies for the needed properties in relation to safety in the literature. However, there are no papers for a safety ontology which includes all the safety properties that are found in ENISA. Furthermore, it is difficult to handle safety and cybersecurity interdependency through the properties of trustworthiness that are found in ENISA. And so, a new ontology for safety is needed to handle this interdependency.

3. NFV safety ontology

This section presents a new safety ontology. This ontology is used by an orchestrator that ensures safety in a NFV framework. Our proposed ontology is able to: (i) describe a variety of different breakdown types related to safety and (ii) help the decision for the best reaction to safety-related anomalies while considering the safety and cybersecurity interdependency. Our ontology-based solution is written in Ontology Web Language (OWL). This is because it provides greater content interpretability, in comparison with eXtensible Markup Language (XML) and Resource Description Framework (RDF). OWL language facilitates the expression of knowledge, and it also provides the means to reason with this knowledge.

As seen in [43], there are many advantages that ontologies bring. These advantages are (i) the modeling clarity which refers to the clear description, (ii) the choice of specificity level which refers to the level of the detailed representation of the content, (iii) the systematicity in information retrieval which makes it possible to access classes and subclasses to get information, (iv) the systematic and coherent definitions where the conceptual information are organized and clarified, and (v) the dynamicity as the ontology is able to represent the concept evolution through time. More specifically, our safety ontology is chosen because of its capabilities to model safety and its properties in a clear way with coherent definitions. Moreover, the possibility to access classes and subclasses in order to retrieve information supports the process of making decisions. These decisions are able to manage the safety and cybersecurity interdependency.

Our safety ontology provides the description of a representation of safety and the properties that it interacts with it according to ENISA. More specifically, all of these properties have the ability to cause anomalies which affect humans and environment. For this reason, the safety principle is decided to be the core of our proposed ontology.

All things considered, a NFV framework consists of a variety of modules. Different orchestrators are able to handle these modules. Thus, for ensuring and handling safety in a NFV framework, an orchestrator is needed. This orchestrator requires a way to identify whether a safety-related anomaly is also affected by cybersecurity. This is where our proposed ontology takes action, since with the help of the reasoner it is able to identify whether an anomaly is both safety and cybersecurity-related. In order to reach to this outcome, the reasoner uses (i) our proposed ontology and (ii) our safety and cybersecurity interdependency rules. NFV safety orchestrator needs this outcome in order to understand whether the mitigation plan creates violations to cybersecurity functionalities during mitigation.

In practice, our ontology consists of three parts: (i) the class of safety, (ii) the concerned elements, and (iii) the object properties (see **Figure 1**).

The first part provides safety as the class of our ontology. The second part describes the elements that are affected by potential safety-related anomalies. These elements are the functionality, structural, operational, and system. Each concerned element is associated with a safety-related anomaly. Thus, these concerned elements are represented as subclasses of safety. Finally, the third part provides the possible

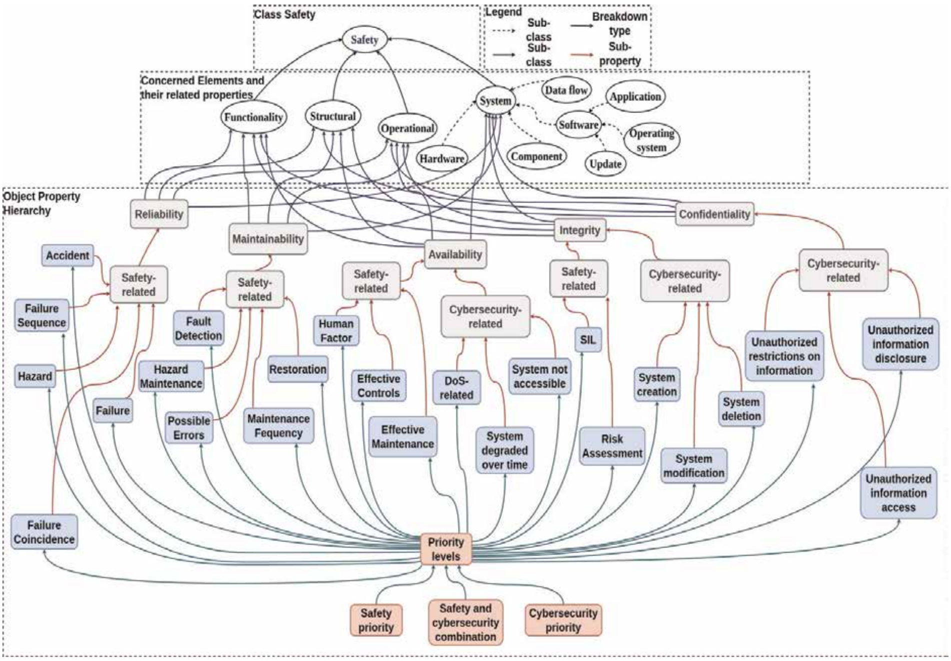


Figure 1.
Safety ontology.

breakdown types related to each one of the safety-related properties as they are found in ENISA. Each one of the safety-related properties and confidentiality is an object property of the concerned elements. And each one of these properties is associated with its relation to safety, cybersecurity, and/or both. By extension, each property is associated with their possible breakdown types. Each breakdown type is then assessed with respect to a priority level which is divided in safety, cybersecurity, and both.

It should be mentioned that our concerned elements and our breakdown types are extracted from the standards: NIST [44], MITER [45], ISO 61508 [46].

In particular, our ontology is able to provide three possible outcomes. The first outcome refers to the anomaly as only safety-related. The second outcome corresponds to an anomaly that is interdependent between safety and cybersecurity. In order to get this specific outcome, it is needed to to describe how the potential anomalies are related to the elements of the ontology. Thus, once a potential anomaly is identified as safety-related, it is then associated with its concerned element. According to the origin property of the anomaly (i.e. reliability, maintainability, availability, integrity, and confidentiality), it is possible to understand if it affects more than one property and if this impact is also creating cybersecurity-related breakdown types. For example, a VNF stops working. This VNF handles the access management of the cybersecurity functions. This event is a reliability problem which affects also availability and integrity. Thus, the reliability-related anomaly is able to affect cybersecurity. In this case, this anomaly is considered as an interdependent one between safety and cybersecurity.

Finally, the third outcome deals with the affected breakdown types, since it is possible to assess the priority level. This priority level indicates that safety, cybersecurity, or both orchestrators can handle the anomaly.

Certain rules are defined in order to handle the safety and cybersecurity interdependency. Based on the rules, the reasoner is able to make decisions for eventually avoiding potential safety and cybersecurity conflicts. There are two types of rules. The first type is composed of three statements concerning: (i) the type of breakdown, (ii) the relation of the anomaly to safety, cybersecurity, or both, (iii) and the affected object property (reliability, availability, maintainability, confidentiality, and integrity-related). According to these rules, it is possible to identify whether cybersecurity is affected through availability and integrity, but also to see how cybersecurity impacts safety through all the safety-related properties. And the second type of rules is composed of two statements considering: (i) the outcome of the first rules and (ii) the breakdown type. Thus, it is both of these types of rules that are used to automate the process of inferring, during decision-making.

4. Interdependency rules

Safety and cybersecurity interdependency rules are the ones that the reasoner uses to understanding whether a safety-related anomaly is also affecting cybersecurity and vice versa. But also, these rules are used to get the indication of which orchestrator between safety and cybersecurity is prioritized for mitigating the anomaly.

4.1 Safety and cybersecurity rules

In general, in order to create these rules, it is taken into account the type of the breakdown and its impact to cybersecurity. More specifically, the type of the breakdown is associated with: (i) the element that is affected by the anomaly and (ii) the specific breakdown type. The impact to cybersecurity refers to the object property that is affected by the specific anomaly. Thus, it is understood that there are three important terms. These terms are (i) the fact that the safety-related anomaly is coming from a cyberattack (*C.A.* from cyberattack) or has the same effect, (ii) the subproperty of the specific object property (*B.T.* from breakdown type), and (iii) the object property that is also affected by the anomaly (*O.P.* from object property). It should be mentioned that our proposed rules are considered only when the event is coming from a cyberattack or has the same effect. Hence, each rule (*I.R.* from interdependency rule) is a set of three statements referring to these terms, with the following form: $I.R. = C.A. + B.T. + O.P.$

As an example, a VNF at a production unit handles the working time scheduling between humans and robots. This VNF is cyberattacked and stops working. This is a cybersecurity anomaly. However, it also impacts safety since humans may be harmed. Thus, it is a safety-related anomaly which comes from the reliability property. This anomaly affects the functionality concerned element, and it can cause the accident breakdown type. It also impacts the availability, confidentiality, and integrity object properties. Thus, in this case, the corresponding rule is $I.R. = C.A. + Accident + All$, where:$All = Integrity + Availability + Confidentiality$.

All things considered, it is understood that each safety-related property has a total number of safety and cybersecurity interdependency rules. To calculate this total number, it is needed to calculate first the total number of our proposed rules for each one of the safety-related properties. In order to make this calculation, we created the following formula: $I.R._{tot\_x} = C.A. \times \sum_{B.T.} \times \sum_{O.P.}$.

In this formula: (i) the $I.R._{tot\_x}$ corresponds the total number of the interdependency rules for each one of the safety-related properties, (ii) the x is substituted by the *re* for reliability, *ma* for maintainability, *in* for integrity, *conf* for confidentiality, and *av* for availability, (iii) the $\sum_{B.T.}$ corresponds to the sum of the subproperties for each one of the safety-related properties, (iv) $\sum_{O.P.}$ refers to the sum of the possible object properties affected for the specific anomaly, and (v) the C. A. is equal to one since it is the Boolean true. It should be mentioned that reliability and maintainability are able to impact cybersecurity. However, availability and integrity are able to affect safety.

For each of the reliability and maintainability, the $\sum_{O.P.}$ is equal to four. This is because these properties can have four different possibilities of impacting cybersecurity. These four different possibilities are through availability, integrity, confidentiality, or all. For integrity, it is possible to impact safety through availability, reliability, or maintainability. Thus, the $\sum_{O.P.}$ is also equal to three. For availability, it is possible to impact safety through integrity, reliability, or maintainability. Hence, the $\sum_{O.P.}$ is also equal to three. Finally for confidentiality, it is possible to affect safety through reliability, maintainability, availability, and integrity. Thus, the $\sum_{O.P.}$ is equal to four.

Thus, the calculated total number of interdependency rules for: (i) reliability is: $I.R._{tot\_re} = 1 \times 5 \times 4 = 20$, (ii) maintainability is: $I.R._{tot\_ma} = 1 \times 5 \times 4 = 20$, (iii) availability is: $I.R._{tot\_av} = 1 \times 6 \times 3 = 18$, (iv) integrity is: $I.R._{tot\_in} = 1 \times 5 \times 3 = 15$, and (v) confidentiality is: $I.R._{tot\_conf} = 1 \times 3 \times 4 = 12$.

The total number of rules to manage the interdependency between safety and cybersecurity is calculated in the following formula. In this formula, the total number of the interdependency rules is the sum of each one of the interdependency rules of the safety-related properties. Thus, the total number of the interdependency rules is

$$I.R._{tot} = I.R._{tot\_re} + I.R._{tot\_in} + I.R._{tot\_av} + I.R._{tot\_ma} + I.R._{tot\_conf}, I.R._{tot} = 85$$

4.2 Priority level rules

This type of rules depends on the outcome of the previous reasoning and rules since it is taken into account the type of the anomaly. Thus, there is only one term for this rule which refers to the related types of the anomaly (*R.T.* from related-type). Each rule (*P.L.* from priority level) is equal to this one term: *P.L* = *R.T.* In case the potential anomaly is safety-related and does not affect cybersecurity, then this rule results that the safety orchestrator has to mitigate this anomaly. In the second case where the safety-related anomaly is impacting cybersecurity, then the outcome of the priority level rule is that both safety and cybersecurity need to handle the anomaly. In the third case where an anomaly is confidentiality and it also affects safety, then the priority level rule decides that only the cybersecurity orchestrator is to handle this anomaly.

The total number of this rule is equal to the sum of the possible related types for an anomaly. The related types for an anomaly are (i) safety-related, (ii) cybersecurity-related, and (iii) safety- and cybersecurity-related. Thus, the total number of the priority level rules is $P.L._{tot} = \sum_{R.T.} = 3$.

An example is provided for better understanding. A VNF that handles the emergency protection of a system stops working. This is identified as a reliability anomaly which is realted to safety. However, in the specifications of our system, this specific VNF is able to cause our system to degrade over time, in case it stops working. The breakdown type of system degraded over time is a cybersecurity-related bone. According to our ontology, this is a breakdown type that is subproperty of availability.

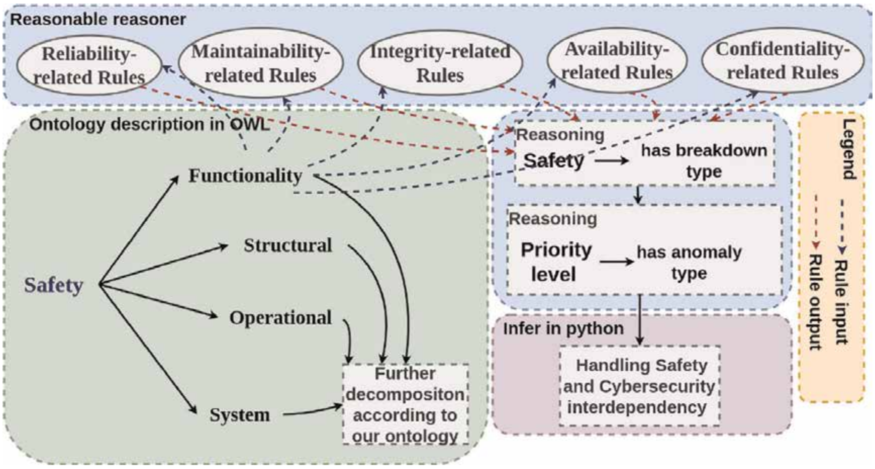

Reasonable reasoner
Reliability-related Rules
Maintainability-related Rules
Integrity-related Rules
Availability-related Rules
Confidentiality-related Rules
Ontology description in OWL
Functionality
Structural
Safety
Operational
System
Further decompositon according to our ontology
Reasoning
Safety
has breakdown type
Reasoning
Priority level
has anomaly type
Infer in python
Handling Safety and Cybersecurity interdependency
Legend
Rule input
Rule output

Figure 2.
Handling safety and cybersecurity interdependency, with the combination of our ontology and a rule-based reasoner.

Thus, it is understood that a safety anomaly has the same effect as a cyberattack to our system, and that both safety and cybersecurity are affected. Moreover, a reliability anomaly has impacted an availability one. Thus, based on the priority level rules, our reasoner decides that the priority-level outcome is for both safety and cybersecurity orchestrators to act upon the anomaly. More specifically, it is suggested that the safety orchestrator can handle the safety-related anomaly, while the cybersecurity orchestrator can handle the cybersecurity-related anomaly.

Considering everything, in **Figure 2**, it is possible to see the functioning model of our solution. Our model consists of three parts which are the ontology, the reasoner, and the outcome. Our ontology is represented in OWL in order to fully described the whole knowledge of safety in one common language. Each property of safety and confidentiality corresponds to a specific set of rules. The reasoner is able to make decisions based on these sets of rules and to provide an outcome that best handles the interdependency between safety and cybersecurity.

5. Evaluation and results

For the implementation and evaluation phase, a testbed is constructed with the intention to test our proposed safety orchestrator. Our testbed is composed of six Virtual Machines (VMs). These VMs are executed in a computer with an Intel core i7 processor which is running at 4.6 GHz. More specifically, the number of threads that are able to execute instructions at once is 16. Our machine uses 15.744 GB of RAM, with an additional Swap memory of 15.6 GB. The OS running is Linux, and more especially Pop OS 20.04 focal which is based on Ubuntu 20.04. Each VM uses 6 GB of RAM and 3 vCPU. Open Source MANO (OSM) handles all the VNF by using the VNF and NS descriptors for instantiation. And Openstack handles the whole architecture of the servers. Furthermore, each orchestrator of our proposed safety architecture corresponds to one VNF and one instantiated VM. It is possible to access these VM via openstack.

Our use case is provided in **Figure 3**. Free5GC includes the functions: (i) Network Repository Function (NRF): serves as a central repository for virtualized functions, (ii) Authentication Server Function (AUSF): supports the authentication of an entity that attempts to access a network, (iii) Access and mobility Management Function (AMF): manages the reachability, registration, mobility, and connection, (iv) Session Management Function (SMF): controls the session, and (v) User Plane Function (UPF): serves for the part of the network that carries the data traffic. In our testbed, the Free5GC core corresponds to the first server and represents the various VNF of a NFV framework. Each one of these functions corresponds to a VNF. Each one of these VNF is able to generate anomalies which are related to virtualized function and service issues with respect to NFV framework. For example, the VNF which corresponds to UPF function is not able to migrate. In this way, it is possible to simulate the issues of a NFV with respect to VNF. This is structured in a docker environment with each VNF occupying a container which uses Ubuntu 20.04.

Furthermore, the safety orchestrator needs to be able to receive the anomaly messages from the rest of the orchestrators. For this to happen, a client–server

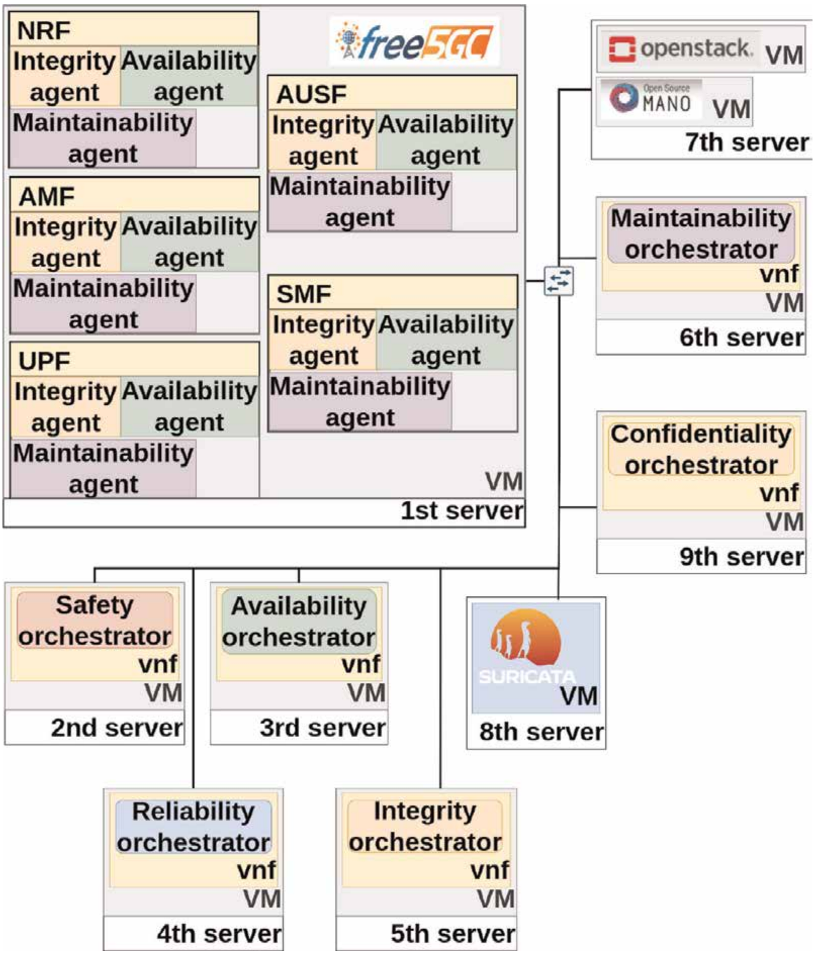


Figure 3.
NFV safety architecture use case.

architecture is implemented. This architecture uses the WSGIserver technology. According to this architecture, (i) safety orchestrator is the server which receives the anomaly message, treats it, and sends back the best response, (ii) and each one of the rest of the orchestrators is the client which sends the anomaly message and then receives the reply. Furthermore, the server sends two types of messages to the client: (i) one message is for sending the course of action to the act module of the client in order to implement them in case that the anomaly type is safety-related, (ii) and the second message is for sending back the anomaly information in case that the anomaly type is not safety-related. In case of a safety-related anomaly, it is possible to handle the interdependency between safety and cybersecurity. This is because the decided course of action is already verified from the reasoner feature in the conflict management module of the safety orchestrator.

The anomaly messages are pulled by a REST API, as it retrieves the message with a GET request. Each anomaly is therefore addressed in the URL, which makes the safety orchestrator able to access and process the message. In this way, the objects are retrieved by this specific URL. At first, they are treated to find the appropriate course of action. And then, they are sent to the Reasonable reasoner in the conflict management module for handling the safety and cybersecurity interdependency. In our case, our implementation handles the anomalies in parallel and more specifically between five processes.

In order for safety to be achieved, it is needed to analyze the results obtained from our implementation. These results are the type of the anomaly and the decided course of action that ensure safety and cybersecurity interdependency. Thus, safety is achieved once the anomaly message is treated and the course of action is decided and sent back to one of the rest of the orchestrators. Moreover, the total number of messages treated shows how effective and stable our proposal is. For this reason, it is important to acquire the time that each message takes to be treated. The shorter the time that an anomaly takes to be treated, the greater the number of the messages are treated. Consequently, the outcomes of our implementation are provided in terms of time which are as follows: (i) the reasoner processing time: the time that the conflict management module takes to decide how to best handle the safety and cybersecurity interdependency, and (ii) the total response time: the amount of time that it takes for one of the orchestrators to receive the response from the safety orchestrator.

For realizing **Figure 4a**, the number of the repeated tests is greater than 30 with 1 minute as a running time per each test. The number of anomalies treated is affected by the time that the reasoner needs to execute the rules in the safety architecture

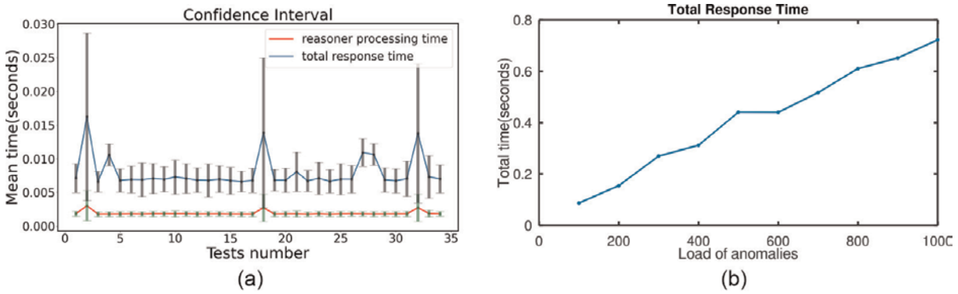


Figure 4.
(a) Reasoner processing time and total response time; (b) total response time's mean value in relation to different load of anomalies per second.

ontology, since it needs to be able to make decisions about the safety and cybersecurity interdependency. **Figure 4a** provides our two metrics. The total response time is in blue, and the reasoner processing time is in red. Each point of the lines corresponds to the mean value of one test. For each mean value point, the above and below standard deviation bars are provided. Some points have greater standard deviation values than others, since standard deviation is affected by the number of the samples and the mean value which both change from test to test. Thus, the points with the higher mean values are the ones with the greatest standard deviation. Finally, our study provides 95% of assurance that an anomaly is treated with confidence interval bounds of 0.0072 s to 0.0088 s for the total response time, and 0.0018 s to 0.0020 s for the reasoner processing time. Consequently, the mean time values for both of our metrics are quite low, and our solution is considered stable. This makes it seem possible to use our solution for real-time systems.

Figure 4b illustrates the total response time. In this test, the number of anomalies that are generated per second are iterated by 100 each time. It is understood that as the load of anomalies increases, the latency of the response time gradually rises. And it is almost linear. This is possible because the anomalies may be generated in 1 second, but the conflict management needs time to treat them all. Moreover, the anomalies are handled in parallel between five processes. However, the latency grows gently which seems to suggest that it is possible to meet the real-time constraints of many applications. Overall, this graph shows that our solution is scalable, and that it can be used for larger architectures.

6. Conclusion

Safety and cybersecurity are able to impact each other in a NFV framework, and so both of them need to be taken into consideration. Thus, it is important to be able to manage safety in a more comprehensive way. But, it is also important to handle the safety and cybersecurity interdependency. In this paper, an ontological-based solution for handling safety is proposed. Moreover, the safety and cybersecurity interdependency rules are proposed. More specifically, our proposed ontology is able to describe safety through the safety-related properties found in ENISA (i.e. reliability, availability, maintainability, and integrity). Together, the ontology and the rules are used by an orchestrator that manages the safety of a NFV framework. This is because the safety orchestrator needs to understand whether a safety-related anomaly is also affecting cybersecurity. This specific information is able to help the safety orchestrator to modify the plan of mitigation in order to avoid and functionality violations between safety and cybersecurity. In order to evaluate and test our solution, a testbed is created. This testbed is a safety and security management in 5G core network. According to the obtained results, our solution is able to ensure safety. Moreover, our solution is scalable, and it can be used in other applications.

Author details

Dionysia Varvarigou[1,2]*, David Espes[1] and Giacomo Bersano[2]

1 Université de Bretagne Occidentale, Paris, France

2 Ikos Consulting, Paris, France

*Address all correspondence to: dionysia.varvarigou@etudiant.univ-brest.fr

References

[1] European Network and Information Security Agency. Measurement Frameworks and Metrics for Resilient Networks and Services: Technical report, Discussion draft. 2011

[2] Nogales B, Silva M, Vidal I, Luís M, Valera F, Sargento S, et al. Using aerial and vehicular NFV infrastructures to agilely create vertical services. Sensors. 2021;**21**:1342

[3] Gruber T. Ontology. 2018

[4] Xing X, Zhong B, Luo H, Li H, Wu H. Ontology for safety risk identification in metro construction. Computers in Industry. 2019;**109**:14-30

[5] Jiang X, Wang S, Wang J, Lyu S, Skitmore M. A decision method for construction safety risk management based on ontology and improved CBR: Example of a subway project. International Journal of Environmental Research and Public Health. 2020;**17**:3928

[6] Single J, Schmidt J, Denecke J. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. Safety Science. 2020;**129**:104747

[7] Fang W, Ma L, Love P, Luo H, Ding L, Zhou A. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. Automation in Construction. 2020;**119**:103310

[8] Zhong B, Li H, Luo H, Zhou J, Fang W, Xing X. Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images. Journal of Construction Engineering and Management. 2020; **146**:04020013

[9] Abdelghany AS, Darwish NR, Hefni HA. An agile methodology for ontology development. International Journal of Intelligent Engineering and Systems. 2019;**12**:170-181

[10] Bento A, Zouaq A, Gagnon M. Ontology matching using convolutional neural networks. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020

[11] Alkahtani M, Choudhary A, De A, Harding J. A decision support system based on ontology and data mining to improve design using warranty data. Computers & Industrial Engineering. 2019;**128**:1027-1039

[12] Ageed ZS, Ibrahim RK, Sadeeq M. Unified ontology implementation of cloud computing for distributed systems. Current Journal of Applied Science and Technology. 2020;**39**:82-97

[13] Kestel P, Kügler P, Zirngibl C, Schleich B, Wartzack S. Ontology-based approach for the provision of simulation knowledge acquired by Data and Text Mining processes. Advanced Engineering Informatics. 2019;**39**:292-305

[14] Banane M, Belangour A. Towards a new scalable big data system semantic web applied on Mobile learning. International Journal of Interactive Mobile Technologies. 2020;**14**

[15] Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS). BS EN 50126. 2017

[16] Frühwirth T, Preindl T, Kastner W. Ontology for rating dependability attributes. In: IECON2022–48th Annual Conference of the IEEE Industrial Electronics Society. IEEE; 2022

[17] Gobin-Rahimbux B. Evaluation metrics for ontology modules. In: 2022 IEEE International Conference on Data Science and Information System (ICDSIS). IEEE; 2022

[18] Lupp D, Hodkiewicz M, Skjæveland MG. Template libraries for industrial asset maintenance: A methodology for scalable and maintainable ontologies. In: Proceedings CEUR Workshop. 2020

[19] Montero JJ, Vingerhoeds R, Grabot B, Schwartz S. An ontology model for maintenance strategy selection and assessment. Journal of Intelligent Manufacturing. 2021:1-19

[20] Iglesias-Molina A., Chaves-Fraga D., Priyatna F., and Corcho O. Enhancing the Maintainability of the Bio2RDF Project Using Declarative Mappings. 2019.

[21] Kilintzis V, Chouvarda I, Beredimas N, Natsiavas P, Maglaveras N. Supporting integrated care with a flexible data management framework built upon Linked Data, HL7 FHIR and ontologies. Journal of Biomedical Informatics. 2019;**94**:103179

[22] Meng K, Cui C, Li H. An ontology framework for pile integrity evaluation based on analytical methodology. IEEE Access. 2020;**8**:72158-72168

[23] Chaves-Fraga D, Ruckhaus E, Priyatna F, Vidal ME, Corcho O. Enhancing virtual ontology based access over tabular data with Morph-CSV. In: Semantic Web. IOS Press; 2021

[24] Nikolaou C, Cuenca GB, Kostylev EV, Kaminski M, Horrocks I. Satisfaction and implication of integrity constraints in ontology-based data access. In: International Joint Conferences on Artificial Intelligence. 2019

[25] Manzoor S, Vateva-Gurova T, Trapero R, Suri N. Threat modeling the cloud: an ontology based approach. In: International Workshop on Information and Operational Technology Security Systems. Springer; 2019;**12**:170-181

[26] Blokland P, Reniers G. An ontological and semantic foundation for safety and security science. Sustainability. 2019;**11**:6024

[27] Pereira DP, Hirata C, Nadjm-Tehrani S. A STAMP-based ontology approach to support safety and security analyses. Journal of Information Security and Applications. 2019;**47**:302-319

[28] Shaaban AM, Schmittner C, Gruber T, Mohamed AB, Quirchmayr G, Schikuta E. Ontology-based model for automotive security verification and validation. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. 2019

[29] Alkhammash E. Formal modelling of OWL ontologies-based requirements for the development of safe and secure smart city systems. Soft Computing. 2020;**24**:11095-11108

[30] Venkata RY, Kamongi P, Kavi K. An Ontology-Driven Framework for Security and Resiliency in Cyber Physical Systems. ICSEA; 2018

[31] Trappey AJC, Trappey CV, Govindarajan UH, Jhuang ACC. Construction and validation of an ontology-based technology function matrix: technology mining of cyber physical system patent portfolios. In: World Patent Information. Elsevier; 2018

[32] Fang Y, Nazila RE, Chimay A. A Knowledge-Based Cyber-Physical System (CPS) Architecture for Informed Decision Making in Construction. In:

Construction Research Congress. American Society of Civil Engineers; 2018

[33] Nuñez DL, Borsato M. OntoProg: An ontology-based model for implementing Prognostics Health Management in mechanical machines. In: Advanced Engineering Informatics. Elsevier; 2018

[34] Polenghi A, Roda I, Macchi M, Pozzetti A. Multi-attribute Ontology-based Criticality Analysis of manufacturing assets for maintenance strategies planning. In: IFAC-PapersOnLine. Elsevier; 2021

[35] Ansari F, Khobreh M, Seidenberg U, Sihn W. A problem-solving ontology for human-centered cyber physical production systems. CIRP Journal of Manufacturing Science and Technology. 2018;**22**:91-106

[36] Sanislav T, Zeadally S, Mois GD, Fouchal H. Reliability, failure detection and prevention in cyber-physical systems (CPSs) with agents. In: Concurrency and Computation: Practice and Experience. Wiley Online Library; 2019

[37] Ali N, Hong JE. Failure detection and prevention for cyber-physical systems using ontology-based knowledge base. In: Computers. Multidisciplinary Digital Publishing Institute; 2018

[38] Venkata RY, Maheshwari R, Kavi K. Simon: Semantic inference model for security in cyber physical systems using ontologies. ICSEA; 2019

[39] Griffor E, Greer C, Wollman D, Burns M. Framework for Cyber-Physical Systems: Volume 1, Overview. Gaithersburg, MD: National Institute of Standards and Technology; 2017

[40] Nguyen TH, Son TC, Bundas M, Balduccini M, Garwood KC, Griffor ER. Reasoning about trustworthiness in Cyber-Physical Systems using ontology-based representation and ASP. In: International Conference on Principles and Practice of Multi-Agent Systems. Springer; 2020

[41] Venkata RY, Brown N, Maheshwari R, Kavi K. A domain-agnostic framework for secure design and validation of CPS systems. International Journal on Advances in Security. 2020

[42] Cho JH, Xu S, Hurley PM, Mackay M, Benjamin T, Beaumont M. Stram: Measuring the Trustworthiness of Computer-based Systems. NY: ACM; 2019

[43] Durán-Muñoz I, Bautista-Zambrana MR. Applying ontologies to terminology: Advantages and disadvantages. Hermes-Journal of Language and Communication in Business. 2013

[44] Ross R, Pillitteri V, Graubart R, Bodeau D, McQuaid R. Developing Cyber Resilient Systems: A Systems Security Engineering Approach. National Institute of Standards and Technology; 2019

[45] Bodeau G. MITRE, Cyber Resiliency Design Principles, Selective Use throughout the Lifecycle and in Conjunction with Related Disciplines. 2017.

[46] International Organization for Standardization. Functional safety of electrical/electronic/programmable electronic safety-related systems. ISO/IEC 61508-7:2010. International Organization for Standardization. 2010

Chapter 4

Climate System Ontology: A Formal Specification of the Complex Climate System

Armita Davarpanah, Hassan A. Babaie and Guanyu Huang

Abstract

Modeling the climate system requires a formal representation of the characteristics of the system elements and the processes that change them. The Climate System Ontology (CSO) represents the semantics of the processes that continuously cause change at component and system levels. The CSO domain ontology logically represents various links that relate the nodes in this complex network. It models changes in the radiative balance caused by human activities and other forcings as solar energy flows through the system. CSO formally expresses various processes, including non-linear feedbacks and cycles, that change the compositional, structural, and behavioral characteristics of system components. By reusing the foundational logic of a set of top- and mid-level ontologies, we have modeled complex concepts such as hydrological cycle, forcing, greenhouse effect, feedback, and climate change in the ontology. This coherent, publicly available ontology can be queried to reveal the input and output of processes that directly impact the system elements and causal chains that bring change to the whole system. Our description of best practices in ontology development and explanation of the logics that underlie the extended upper-level ontologies help climate scientists to design interoperable domain and application ontologies, and share and reuse semantically rich climate data.

Keywords: climate system, ontology, complex system, climate change, climate data

1. Introduction

The solar powered and highly complex climate system consists of five interacting subsystems of atmosphere, hydrosphere, lithosphere, cryosphere, and biosphere [1]. The system has evolved to maintain its radiative balance through self-organization [2, 3] and adaptation [4] over time. Radiative perturbations caused by natural external forcings such as changes in the solar cycles and volcanic eruptions have periodically changed the climate system over long temporal and spatial scales [5]. The system has adapted to deal with the continuous input of naturally produced carbon dioxide and volcanic aerosol to its atmosphere and oceans, by storing carbon in plants and carbonate sedimentary rocks and coral reefs [6].

More recently, human activities such as burning of fossil fuel and deforestation are increasing the levels of carbon dioxide in the atmosphere and oceans [7, 8]. Increased concentrations of CO_2 and other greenhouse gases are disturbing system's radiative balance through enhanced greenhouse effect which leads to warming of the atmosphere beyond natural balanced levels [9]. The complex climate system is responding to these changes by nonlinear feedback processes and modifying the interactions among its internal components [10]. Through self-organization, the system brings change in its established network structure and climates. Climate change, for instance, brings change to the spatial and temporal pattern, frequency, and intensity of extreme events in system components, affecting all kinds of ecosystems, including social system [11].

The knowledge about the complex interactions in the climate system is well known in the climatology community (e.g., [12–14]). Recently, the Intergovernmental Panel on Climate Change (IPCC) has advanced the knowledge by emphasizing the effects of anthropogenic activities in the climate system [5, 15]. The IPCC reports, however, are written in a natural language. As such, the knowledge cannot be understood by software without natural language processing (NLP). The vast knowledge described in these and other reports about the interactions in the climate system is not easily accessible or understandable by decision and policymakers who are expected to prescribe plans to mitigate climate change impacts. One way to make the knowledge more accessible to software is to develop ontologies that translate the known facts, expressed in natural language, into the logic-based, machine-processible Web Ontology Language (OWL) [16]. Given the complexity of the climate system in which all sorts of nonlinear physical, chemical, and biological processes occur, only ontologies that are based on robust description logic [17] of well-established upper ontologies can reliably model these complex relationships. To this end, we have developed the Climate System Ontology (CSO) based on the widely used, top-level Basic Formal Ontology (BFO) [18, 19] and mid-level Common Core Ontologies (CCO) [20–22], Relation Ontology [23], and Phenotype And Trait Ontology (PATO) [24]. Our Climate System Ontology semantically models the system variables (e.g., concentration of CO_2 in the atmosphere or ocean) and core interactive processes among the climate system's components (e.g., change in oceanic circulation, retreat of glaciers, and atmosphere-ocean heat exchange). It explicitly formalizes the impact of anthropogenic activities (e.g., fossil fuel emission, emission of greenhouse gases, and land use) on climate variation. The semantic model represents the natural greenhouse and other effects by representing the influences of trace gases (e.g., greenhouse gases), aerosols, clouds, and other entities on system's energy balance. The CSO ontology also models major feedback and energy transfer processes and negative and positive variations in the radiative forcing.

In this chapter, we provide a detailed description of the best practices that we have applied to develop the Climate System Ontology, CSO, by extending upper ontologies. We also explain how the description logic, inherited from the reused upper-level ontologies, enriched the CSO knowledge model and enabled it to model complex concepts in the climate system. By being publicly available in the cloud-based GitHub software repository (GitHub - adavarpa/Climate-System-Ontology-CSO: An ontology for Climate system), our Climate System Ontology allows climate scientists to build their own domain and application ontologies that model, for instance, ice core, ocean water composition, permafrost melting, or extreme events. The CSO ontology enables integration and federation of heterogeneous data and facilitates search, retrieval, and analysis of climate data.

This chapter is structured as follows: Section 1 introduces the problem and the objectives of making the Climate System Ontology (CSO). Section 2 presents a concise description of the climate system from a complex system perspective. Section 3 introduces the methodology and the foundational logics and structures of the upper ontologies which were used to construct CSO. Section 4 presents the results of our modeling of parts of the climate system. Section 5 discusses the advantages of developing the CSO domain ontology by extending the upper ontologies and presents models of the most complex processes in the climate system. It also explains how the ontology, currently available on Github (see above), can be applied by domain scientists. This is followed by Section 6 which summarizes our work.

2. Climate system

2.1 Flow of energy in the climate system

Logical modeling of the network structure and dynamics of the climate system in an ontology requires a formal, explicit specification of our conceptualization of the interactions among the components of this system [25]. In this section, we provide the background knowledge of the climate system based on the descriptions presented by the IPCC reports [5, 15, 26, 27], and the reports by the Royal Society and the U.S. National Academy of Sciences [11, 28]. We then model this knowledge in the next sections.

The open atmospheric, hydrospheric, cryospheric, lithospheric (land surface), and biospheric components of the climate system continuously interact through many physical, chemical, and biological processes over a wide range of spatial and temporal scales [26]. The climate system is driven by solar radiation, mostly in the visible short-wave and near-infrared, and, to some extent, in the ultraviolet part of the electromagnetic spectrum [29, 30]. About a third of this incoming solar radiation is reflected back into space by clouds, the atmosphere, and land surface [31]. The rest is either absorbed by the atmosphere or used to heat the land and oceans. After absorbing part of incoming radiation, the warmed land surface returns the energy as heat (infrared radiation) and water vapor to the atmosphere. Heat, carried through atmospheric circulations by water vapor, is released to the atmosphere through condensation [26]. The infrared radiation emitted from Earth's surface is partly absorbed by greenhouse gases and clouds in the atmosphere [32]. The greenhouse gases and clouds re-emit the absorbed heat in all directions, leading to its entrapment, and warming of the lower atmosphere and Earth's surface. This is the natural greenhouse effect which is a part of Earth's energy balance. Reflection of the incoming radiation by clouds more than compensates for their warming effect, leading to a net cooling of the system [33].

The average net radiation at the tropopause, which for the sake of energy calculations is assigned as the top of the atmosphere, is zero [34]. The net radiation can change due to a change in the incoming solar radiation or the emitted infrared radiation [35]. The imbalance caused by these variations is called radiative forcing [27]. External natural changes such as solar cycles and explosive volcanic eruptions that eject aerosols into the atmosphere bring variation in the radiative forcing [36]. Positive or negative radiative forcings lead to an average increase or decrease of surface temperatures, respectively [33]. In both of these cases, the system needs to restore the radiative balance. Internal processes and feedbacks [37] in the climate system also cause radiative imbalance by affecting the reflected solar radiation and emitted infrared radiation [26].

The positive forcings are induced by changes in the composition of system components, such as increased greenhouse gas and aerosol concentrations in the atmosphere and oceans, for example, due to combustion of fossil fuels through human activities (e.g., [38]). These changes lead to higher surface-tropospheric and sea water temperatures, along with increased acidification of sea water that affects the carbon sink in corals [39, 40]. These, and other processes (e.g., biomass burning, emission of chlorofluorocarbons that destroy stratospheric ozone layer), cause anthropogenic perturbation of the radiative balance in the system that impact climate as the system tries to restore the balance [41]. The increased concentration of anthropogenic greenhouse gases in the atmosphere leads to a decrease in the amount of heat that is lost to space at higher altitudes, causing a positive radiative forcing [26]. This is called enhanced greenhouse effect [42, 43]. The climate system responds to the radiative imbalance through its various complex internal processes and feedbacks that lead to climate change [15].

Components in the climate system respond to the internal variability and forcings through nonlinear feedback cycles that are essential features of all complex systems [44] (see next section). Due to their different physical properties (e.g., heat capacity and thermal conductivity), components of the climate system have different response times to variations brought by external forcings [45]. This nonlinear feedback mechanisms and other internal interactions among system components keep the climate system in a constantly varying state characterized with large-scale climate variability such as the El Niño-Southern Oscillation, North Atlantic Oscillation, North-South dipole structure, and Antarctic Oscillation which occur due to ocean–atmosphere interaction (e.g., Rodríguez-Fonseca; [46]). Climate variability can result internally due to the natural interactions among system components or externally by radiative forcings (e.g., [47]). The occurrence and intensity of certain low probability, extreme weather events such as heat wave, drought, and flooding correlate with climate variability [48]. Human activities such as agricultural practices, forestry, and land use lead to anthropogenic forcings that cause variations in the climate (e.g., [38]). For example, emission of methane and nitrous oxide gases through agricultural and industrial practices increases the tropospheric ozone (a greenhouse gas) (e.g., [49]).

2.2 Complex system perspective of the climate system

The climate system is an adaptive, dissipative system of a network of numerous independent, nonlinearly interacting nodes (components and elements). The system, defined by the diversity, adaptability, connectedness, and mutual dependency of its heterogeneous components, continuously interacts with its environment (space) through the transfer of solar energy. The interactions among the system components change the state of the whole system as it adapts to internal changes and external perturbations [50, 51] such as the positive radiative forcings caused by solar cycles and human activities.

As a complex system mostly stays in the subcritical, far-from-equilibrium state of the 'edge of chaos' between a stable (low complexity) state and unstable state of chaos [52]. When the system is driven far from equilibrium, for example, through positive radiative forcings caused by human activities, it reaches a threshold of instability (critical level). The transition between the states of the 'edge of chaos' and chaos leads to repeated phase changes (e.g., evaporation and condensation) and cascades (e.g., occurrence of extreme climate events) [53].

The cascades are followed by a return to the slow relaxation stage (subcritical state) to repeat the process (e.g., [54]). The subcritical state [55] is characterized by continuous natural processes (e.g., incoming solar radiation, reflection of incoming radiation, and emission of land surface infrared radiation to atmosphere) and co-evolution [56, 57]. Changes brought by internal and external forcings in one component (e.g., atmosphere and land surface) continuously transfer to other components and reconfigure the network of links among the components [51]. This may lead to co-evolution which occurs when system components simultaneously change due to their interdependency and mutual adaptation [4]. An example of co-evolution in the climate system is when the atmosphere and oceans both become warmer due to the atmosphere–ocean exchange of heat as a result of increased concentration of greenhouse gases in the atmosphere. Another example is an increase of CO_2 in the atmosphere and simultaneous decrease of pH in the ocean water as more CO_2 is absorbed by the oceans [58, 59].

At the critical points, internal and external forcings at the constituent (micro) level lead to spontaneous emergence of order at the whole system (macro) level by the appearance of new properties, random and unpredictable behavior, structure (network of links between nodes), and pattern. The order appears at the macro-level as a result of nonlinear interactions at the micro-level (components) through self-organization [2, 3]. The emerged properties at the macro-level affect those at the micro-level [60, 61]. Since the system is decentralized, a failure at the micro-level does not bring a failure at the whole system level [61].

The self-organization at the critical points leads to the autonomous formation of a preferred configuration (attractor) through nonlinear feedbacks that better conforms (adapts) to the changing environment. The new pattern (e.g., a climate pattern) brings more effective coordination and cooperation among the system elements [61] by reordering the composition and relationships (links) among the system components (nodes) and even creating new ones (e.g., new circulations in ocean; more frequent adverse climate events) [62]. The new self-organized structure, maintained through continuous flux of energy (e.g., through increased anthropogenic radiative forcing), promotes specific behavior, such as climate change, in the system [3]. Self-organization is a dynamical and adaptive feature of the climate system that allows it to acquire and maintain spatial, temporal, or functional structure that leads to increased order [60]. The structure brought by self-organization is maintained through a constant source of energy (e.g., solar radiation) that allows the system to adapt to dynamic changes (e.g., warming of the atmosphere) through a variety of behaviors (e.g., negative feedbacks) allowing the behaviors to restrict to a small part of its state space (i.e., around the attractor) [56, 63]. The emerged self-organization and nonlinear processes that occur during the unstable chaos state are scale-invariant, governed by power laws [50, 64], and produce emergent variations in the system over a wide range of scales. The critical points themselves evolve over time as driving forcings change.

The order produced by the formation of new patterns and structures, through self-organization, is the product of non-equilibrium in the far-from-equilibrium climate system [65, 66]. These patterns (e.g., of climate) are the result of the interaction of the system with its environment (space outside of atmosphere) [62] through input and output of energy. By continuously getting input, such a dissipative and adaptive system can achieve dynamic equilibrium while still doing work, recycling mass (gas, aerosol), and transforming different forms of energy (radiation).

A change such as an increased water vapor content in the atmosphere does not remain proportional to its causal process (e.g., evaporation of ocean water) for long because of the feedback loops that redirect the output of the process (water vapor) back to the original process as input through intermediary processes (e.g., amplification of temperature due to higher water vapor content in the atmosphere). This feedback cycle strengthens or weakens the output of the original change process (evaporation), causing a larger change (positive feedback) or a reduced or eliminated change that brings the system back to equilibrium (negative feedback) [67]. In other words, the outcome of a process is necessary for the process to proceed in a positive feedback (a form of self-cause) [68]. Negative feedbacks dampen the original process and bring stability to the system. Positive and negative feedbacks maintain the emerged self-organized forms [69] within and across system levels [70]. Thus, feedbacks, as nonlinear, recursive processes, probabilistically lead to adaptive or chaotic outcomes or an equilibrium state [69] and result in emergent properties that are absent in the system or its components [71].

3. Methodology

In this section, we describe the structure of the imported upper ontologies and their resources for modeling the climate system. We chose the Common Core Ontologies (CCO) [20–22] to develop our Climate System Ontology (CSO) because these mid-level ontologies extend the logical foundation of the Basic Formal Ontology [18, 19], which is a simple, standard top-level ontology with an extensive scientific user base [72, 73]. The CCO set adds many useful classes to the BFO class hierarchy and introduces several new object properties (through the Extended Relation Ontology) in addition to the ones defined by the Relation Ontology [23] that are available in BFO.

To make it easy for readers to distinguish the CSO class names from names that are defined in the imported upper-level ontologies, we adhere to the following naming scheme throughout this chapter. Imported BFO, CCO, and PATO class names are preceded with their namespace prefix (e.g., bfo: process, cco: Change, pato: quality). Moreover, the BFO and PATO class names begin with a lower-case letter (bfo: material entity, pato: variability) compared to the CCO classes that begin with a capital letter (e.g., cco: Statis, cco: Temperature). Throughout this chapter, we format the CSO class names with Small caps font and capitalize the first letter of each word (e.g., Ocean, Positive Feedback, Snow-covered Surface, and Global Mean Sea level Change).

The CCO set was downloaded as the 'CommonCoreOntologies-master' zip folder from the GitHub [74] cloud-based repository, which was then uncompressed and saved in a working directory. PATO was also downloaded and placed in the CommonCoreOntologies-master uncompressed folder as pato.owl. To access and reuse classes in the upper-level ontologies, our CSO ontology (cso.owl) directly imported PATO and the 'MergedAllCoreOntology-v1.3.ttl' file from the 'cco-merged' folder in the CommonCoreOntologies-master directory. However, since the extracted CCO-master package already included the bfo.owl and ro.owl files in its 'imports' folder, there was no need to directly import these ontologies in the cso.owl file. The Climate System Ontology (cso.owl) was then built using the Protégé editor [75, 76] and placed in the same CommonCoreOntologies-master folder that contained the individual CCO turtle (.ttl) files and pato.owl.

As a best practice [77], we adhered to the 'single inheritance rule' and designed each class in the Climate System Ontology as a subclass of only one imported BFO, CCO, or PATO class. We also reused the object properties that were available in CCO and RO to relate class instances in the CSO. The CSO classes were later expanded to include their necessary and/or sufficient characteristics by constructing compound logical statements (axioms) (see below). To better understand the design of the CSO domain ontology, we evaluate the logics that underlie the class hierarchy of the imported upper ontologies in the remaining part of this section. All examples throughout this chapter are from the Climate System Ontology (CSO).

The Basic Formal Ontology (BFO) [18] classifies all entities (e.g., things that exist or operate in the climate system) as either bfo: continuant or bfo: occurrent. Continuants persist as whole entities in time. These entities have material and immaterial parts but do not have any temporal part. The bfo: continuant class includes the bfo: generically dependent continuant, bfo: independent continuant, and bfo: specifically dependent continuant subclasses. CCO adds the cco: Information Content Entity class under the bfo: generically dependent continuant to represent class of entities whose instances are information content for an Information Bearing Entity. For example, a plot of the global mean surface temperatures vs. time, a table that lists these data, and a report that describes the data, are instances of the cco: Information Bearing Entity that 'carry' the same cco: Information Content Entity (i.e., global mean surface temperature) in different ways. At any time, t, a bfo: generically dependent continuant 'generically depends' on another entity (i.e., Information Bearing Entity). For example, the 'global mean surface temperature' information content 'generically depends' on its carriers (the plot, table, or report).

CCO defines three subclasses of the 'Information Content Entity' class: 'Descriptive Information Content Entity', Designative Information Content Entity', and 'Directive Information Content Entity' that are used for data and information modeling [21, 22]. The 'Descriptive Information Content Entity' consists of propositions that 'describe' some entity and includes the 'Measurement Information Content Entity' (describes extent, dimensions, quantity, or quality of an entity), 'Measurement Unit' (describes a magnitude of a physical quantity), 'Predictive Information Content Entity' (describes an uncertain future event), 'Reference System' (describes a set of standards for organizing data), and 'Representational Information Content Entity' (consists of a set of propositions or the content of an image that represents some entity, e.g., a Satellite Image of a Hurricane). The 'Designative Information Content Entity' consists of a set of symbols that 'designate' or denote some entity. This allows modeling identifiers, abbreviated names such as acronyms (e.g., ENSO and NAO that designate El Niño-Southern Oscillation and North Atlantic Oscillation, respectively), and chemical formulae (CH_4, CO_2). The 'Directive Information Content Entity' consists of propositions or images that 'prescribe' some entity. It allows modeling concepts such as National Climate Change Strategy, Policy for Climate Change Adaptation, Energy Security Goal, and Climate Model.

A bfo: independent continuant includes the bfo: immaterial entity and bfo: material entity. The immaterial entity class includes boundaries (under bfo: continuant fiat boundary) and sites (e.g., the eye of a hurricane) which can change location. Boundaries demarcate material entities, e.g., Sea level, Sea Surface, and Vegetation-covered Surface. A bfo: site is a 3D immaterial entity bounded by material entity, such as a Polar Environment, Soil Environment, a Region of High Salinity, Wet Tropical Region, a cco: Country, or a cco: City. A material entity has portion of matter as part and is a spatially extended independent continuant that maintains its identity through time even when gaining or losing parts. The bfo: material entity allowed us to build classes in CSO such as Climate System Component (e.g., Hydrosphere and Cryosphere), Coast, Ice Field, Atmospheric Layer (e.g., Troposphere), Water Body (e.g., Ocean and Lake), Fossil Fuel, Aerosol, Greenhouse Gas (e.g., Water Vapor), Sample, and Water. CCO provides many classes under its Artifact class (a subclass of bfo: object) that allowed modeling concepts such as Energy-related Carbon Source Facility, Storage of Carbon, Green Infrastructure, Buoy, Drifter, Satellite Sensor, and Ship. A bfo: object aggregate is a group of objects that can be partitioned into mutually exhaustive and pairwise disjoint objects [19]. The object aggregate class let us model the CSO classes of Global Community, Species, Civil Society, Climate System, Forest, and Grassland Ecosystem.

The bfo: specifically dependent continuant inheres in (i.e., is borne by) an independent continuant (the bearer) by virtue of how the bearer is related to other entities [19]. It includes the bfo: quality and bfo: realizable entity subclasses. Examples of quality in CSO are: Amount of Ice, Climate Change Benefit, Capacity for Adaptation, Carbon Intensity, Air Quality, Humidity, Lake Area, and Precipitation Deficit. PATO adds pato: quality in addition to the bfo: quality class. The pato: decreased quality class was used, for example, to model CSO classes of Deceased Amount of Emitted Infrared Radiation from Earth Surface, Decreased Extent of Arctic Sea Ice, Decreased pH of Sea Water, and Decreased Ocean Water Salinity. The pato: increased quality class allowed modeling CSO classes such as Increased Aerosol Content, Increased Net Energy in the Climate System, Increased Intensity of Drought, Increased Ocean Acidity, and Increased Concentration of Carbon Dioxide. PATO also provides the physical object quality, process quality, qualitative quality, and variability classes. These PATO classes allowed modeling many of the qualities of the climate system components in CSO such as the Concentration of Carbon Dioxide, Extent of Arctic Sea Ice, Precipitation Pattern, Glacier Volume, Ocean Water Composition, and Thermal Conductivity.

The bfo: realizable entity is a bfo: specifically dependent continuant that inheres in a bfo: independent continuant. Instances of realizable entities are realized in specific processes. Realizable entities include bfo: disposition (e.g., cco: Color, cco: Albedo, Cosmic Ray Shielding Disposition, Resilience, Risk, Security, Climate Vulnerability) and its bfo: function subclass (e.g., Sensor Artifact Function, and Ecosystem Function), and bfo: role (e.g., Policy Making Role, Driver of Climate Change Role, Positive Radiative Forcing Role, Greenhouse Gas Role, and Proxy Role).

The bfo: occurrent class and its underlying CCO subclasses provide a wide range of mechanisms to model dynamic aspects of the climate system. BFO defines the process, process profile, process boundary, spatiotemporal region, and temporal region classes. CCO adds several classes to each of the BFO classes, increasing their potential for modeling the climate system. These include cco: Act and its cco: Intentional Act subclass, cco: Change, cco: Effect, cco: Natural Process, and cco: Stasis. The cco: Act is a process in which an agent (e.g., a human or group of people) plays a causative role. The cco: Act class is used in CSO to define anthropogenic activities such as Emission of Greenhouse Gases, Forestry, Irrigation, and Land use. The cco: change class allows defining change in the climate system (Climate Change, Forcing, Change in Net Radiation, and Change in the Radiative Balance), in its cycles (Change in Water Cycle), its processes (e.g., Change in Atmospheric Circulation), and in component qualities (e.g., Change in Atmospheric Pressure, Change in Temperature, Change in Humidity, Change in Albedo, and Change in Infrared Radiation). It also provides classes that enable modeling a decrease or increase in a generically or specifically dependent continuant. For example, it allowed CSO to model Decrease in the Extent of Arctic Sea Ice, Decrease in Ocean water Salinity, and Decrease in Temperature of Earth Surface. It also allowed modeling Increase in CH_4 level, Increase in Atmospheric Opacity, and Increase in Background Surface Ozone. The cco: Change also defines ways to model loss or gain of dependent continuants, e.g., loss of quality (Loss of Ice Sheet Mass), loss of disposition (Loss of Well-being, Loss of Health), and loss of function (Loss of Ecosystem Function).

The cco: Effect class enabled CSO to define subclasses for Adverse Effect (e.g., Adverse Economic Effect and Adverse Human Effect) and Climate Change Impact (e.g., Impact on Ecosystem, Impact on Infrastructure; Impact on Species). The cco: Natural Process allowed CSO to model solar and other processes in the component of the climate system, e.g., in the atmosphere (e.g., Filtering of Solar Ultra-violet Radiation, Cloud Formation, Precipitation, and Wind), in oceans (e.g., Heat Transfer, Ocean Current, and Upwelling), in the biosphere (Evapotranspiration, Photosynthesis, and Respiration), and in the cryosphere (Flowing of Outlet Glaciers, Melting, and Retreat of Glaciers). The cco: Stasis class was useful for modeling processes, such as drought and glaciation, through which some independent continuants endure in an unchanging condition over a period of time. For example, the 20th Century Warming, Little Ice Age Cooling, and Drought were modeled under the cco: Stasis of Quality. The bfo: temporal region class defines zero- and one-dimensional temporal regions that allowed us to model concepts such as Glacial Period, Interglacial Period, Ice Age, Period of Abnormally Dry Weather, Winter, and Season (e.g., Percolation Season and Runoff Season) in CSO.

4. Results

4.1 Semantic modeling

Ontologies such as CSO consist of a controlled vocabulary modeled by named and defined classes that represent concepts in the domain knowledge (e.g., climate system). They model different kinds of relations among the individuals that are instances of these classes. These knowledge models are built at different levels based on their generality. Top- and mid-level ontologies define general concepts and are designed to be extended by domain ontologies [78]. A domain ontology is a formal (i.e., logical), explicit specification of conceptualizations in a specific area of interest (e.g., climate system) [25]. As a domain ontology, the Climate System Ontology (CSO) must include classes that formally represent concepts in the climate system such as radiation, global warming, atmosphere, adverse climate impact, and relations that are known to exist among their instances.

A class (e.g., Ice Sheet, Ocean, or Island) in the ontology describes the type for its instances (i.e., an ice sheet, ocean, or island). Class descriptions are commonly complex because they must represent the complete set of characteristics of the concept that they represent. Axioms are often needed to fully describe the complete characteristics of a class. In protégé, axioms are built using the 'SubClass Of' or 'Equivalent To' options in the class description panel applying different logical constructs. The 'Equivalent To' option allows defining both necessary and sufficient conditions for a class 'in one logical statement' using the logical 'and' and 'or'. The 'SubClass Of' option allows defining only the necessary conditions 'in one or more, separate logical statements'. As an example, the full description of the Ice Sheet class is expressed in the following paragraph.

Instances of the Ice Sheet class have several characteristics that must be represented in the class description. By extension, an Ice Sheet is a Cryospheric Object that is made of thick Ice and has a continent-scale extent. It also participates (as input) in several processes such as 'flowing outward from a high central gently-sloping ice plateau' and 'storage of a large amount of water'. These complex descriptions require building several axioms for Ice Sheet, applying the 'SubClass Of' option (**Figure 1**). It also requires making the bfo: continuant classes of Ice (a subclass of Cryospheric Material, placed under bfo: object), Continental Width (under pato: size, a subclass of pato: morphology, under pato: quality), Ice Sheet Thickness (under pato: thickness, a subclass of pato: size), and bfo: occurrent classes of 'Flowing Outward from a High Central Gently-sloping Ice Plateau' and 'Storage of a Large Amount of Water' under the Cryospheric Process class (a subclass of Natural Internal Process, under cco: Natural Process). The first Cryospheric Process also includes sub-processes (e.g., Flowing of Ice Streams and Flowing of Outlet Glaciers) through the 'cco: has process part' object property. All of these natural processes that occur in some Cryospheric

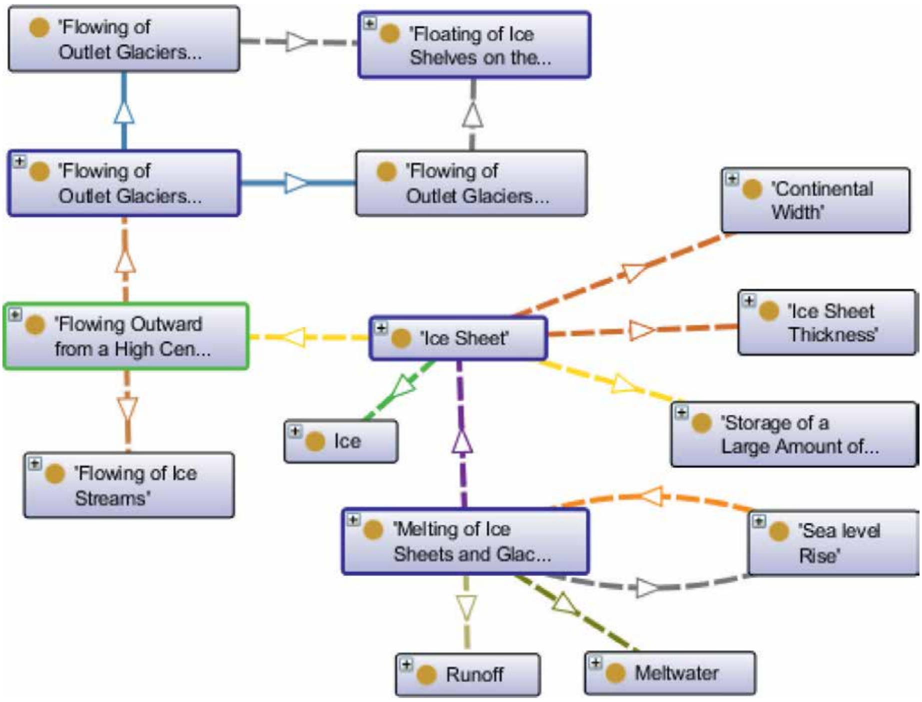

'Flowing of Outlet Glaciers...
'Floating of Ice Shelves on the...
'Flowing of Outlet Glaciers...
'Flowing of Outlet Glaciers...
'Continental Width'
'Flowing Outward from a High Cen...
'Ice Sheet'
'Ice Sheet Thickness'
Ice
'Storage of a Large Amount of...
'Flowing of Ice Streams'
'Melting of Ice Sheets and Glac...
'Sea level Rise'
Runoff
Meltwater

Figure 1.
Modeling the Ice Sheet class using imported classes and object properties of CCO, BFO, and PATO ontologies. Ice sheet has quality Continental Width and Ice Sheet Thickness. Ice Sheet is input of some Flowing Outward from a High Central Gently-sloping Ice Plateau. Flowing Outward from a High Central Gently-sloping Ice Plateau has process part Flowing of Ice Stream and has process part Flowing of Outlet Glaciers. Flowing of Outlet Glaciers has subclass Flowing of Outlet Glaciers into Ice Shelves and Flowing of Outlet Glaciers into the Sea. Flowing of Outlet Glaciers into Ice Shelves process starts Floating of Ice Shelves on the Sea. Flowing of Outlet Glaciers into the Sea process starts Floating of Ice Shelves on the Sea. Ice Sheet is input of some Storage of a Large Amount of Water. Ice Sheet is made of Ice. Melting of Ice Sheets and Glaciers has input Ice Sheet, has output Meltwater, is cause of Runoff, and process starts (i.e., is cause of) Sea level Rise. Sea level Rise process preceded by Melting of Ice Sheets and Glaciers. In all diagrams in this chapter, dashed lines are relations that are represented by the object properties and point from a property's domain (subject) class to its range (object) class (see below for explanation). Each solid arrow represents the has subclass relation and points from a class to its subclass. The diagram was made by the OntoGraf plugin in Protégé.

object (i.e., Ice) occur during a bfo: one-dimensional temporal region (e.g., a period of time). Ice sheets also participate (as input) in the Melting of Ice Sheets and Glaciers class, with Meltwater as output. Melting also causes other processes, such as Runoff and Sea Level Rise, to occur. The Sea Level Rise process leads to (i.e., has output) Increased Sea Level. These processes and changes are shown in **Figure 1**.

In the Climate System Ontology, relations among class instances are modeled through the RO and CCO object properties. Each object property relates instances of a domain class to instances of a range class (through the rdfs: domain and rdfs: range constructs) [79]. For example, in the statement: 'Class-A *relates-to* Class-B', Class-A is the domain, and Class-B is the range, for the *relates-to* object property. Each CCO or RO object property also has other built-in properties and meta-properties such as owl: disjointWith, owl: inverseOf, owl: FunctionalProperty, and owl: TransitiveProperty [16] that provide additional logic for relationships.

Knowledge is the sum of the facts that are known to be true in the domain of discourse in some point in time. For example, the two statements: 'carbon dioxide is a greenhouse gas' and 'meltwater is a product of melting' are known to be true in climate science. An ontology, as a model of a knowledge in a field of study (e.g., climate system), is developed by building numerous logical statements that represent such known facts. Each of these formal statements has three parts: a subject (S), a predicate (P), and an object (O). In OWL, properties stand for the predicates. The 'SPO triples' are the building blocks of knowledge representation. Ontologies are developed by modeling known facts from knowledge repositories (e.g., books, papers, and reports) in the domain, and defining triple SPO statements in logical 'axioms' (e.g., Melting of Ice Sheet cco: *process starts* Sea level Rise). To enhance reading, the reader may ignore the namespace prefixes in such triple statements. Doing so, the above statement is simply read as: 'melting of ice sheet process starts (i.e., is cause of) sea level rise'. The fact that emission of halocarbons leads to stratospheric ozone depletion and to positive radiative forcing can explicitly be expressed by the following two SPO statements: 'Emission of Halocarbon cco: *is cause of* some Stratospheric Ozone Depletion', and 'Emission of Halocarbon cco: *is cause of* some Positive Radiative Forcing'.

The Protégé editor applies OWL 2 [80] which is based on description logic [17]. By extending BFO, CCO, PATO, and RO, the Climate System Ontology inherits the foundational description logic that underlies these imported upper ontologies. CCO and RO object properties explicitly define the type of domain and range classes for each object property. The built-in description logic of these ontologies guarantees the initial consistency and coherency of the CSO domain ontology. As a good practice and to save time during debugging, we continuously ran the HermiT 1.4.3.456 reasoner [81–85], in Protégé, after each major change to the ontology, for example, after adding a new axiom. This assured consistency and coherency of the ontology.

4.2 Climate system ontology

In this section, we describe the construction of the Climate System Ontology (CSO) based on the logical foundations of the imported CCO, BFO, PATO, and RO ontologies which were described above. From an ontological perspective, each of the main components of the climate system (e.g., Hydrosphere, Atmosphere) is a bfo: fiat object part (a subclass of bfo: material entity), associated with theoretically drawn divisions. Fiat boundaries do not coincide with physical discontinuities. These material fiat parts are demarcated by a bfo: two-dimensional fiat boundary which demarcate material entities (e.g., the Equator and Global Mean Sea Level) or immaterial entities (surfaces of cave chambers and boundary of the ozone hole). The location of these two- or three-dimensional fiat boundaries (e.g., Land Surface and Snow-covered Surface) are defined relative to material entities (Rock, Snow, and Vegetation). The CSO examples of fiat boundaries include the cco: Sea Level between the Atmosphere and Hydrosphere, the Land Surface between Atmosphere and Lithosphere, and the Snow-covered Surface between the Cryosphere and Atmosphere. The fiat object parts have their own parts. For example, the Atmosphere has Troposphere, Stratosphere, and other types of Atmospheric Layer as fiat object parts; the Lithosphere has the Northern Hemisphere, Southern Hemisphere, and Mid-latitudes; Ocean has fiat layers such as 'Upper Ocean', 'Top Centimeter Skin', and 'Top Few Meters' as part. These parts and sub-parts consist of different kinds of material objects. The material parts such as Ocean, Land, and Permafrost extend over 3D space and have portion of matter among their proper and improper continuant sub-parts. For example, a molecule of Oxygen in the Atmosphere consists of oxygen atoms, and Soil is made of Mineral, Water, Organic Matter, and Air. Some objects such as Ship, Buoy, Sensor, Storage Facility, and Water Treatment Facility are categorized as cco: Artifact. Other parts are of the bfo: object aggregate types, consisting of disjoint parts that can lose or gain parts while maintaining their identity. A good example is the Climate System which can gain material (Anthropogenic Greenhouse Gas and Volcanic Aerosol) or lose mass and energy in its parts (e.g., Melting of Glaciers and Outgoing Radiation).

The bfo: material entities such as Glacier, Groundwater, Aerosol, and Ozone have characteristics that can be categorized under bfo: quality or pato: quality. For example, the Atmosphere has attributes (qualities) such as Aerosol Content, Heat Content in the Atmosphere, Temperature, and Concentration of Greenhouse Gas; Soil has Soil Moisture Content; Glacier Ice has pato: age; Sea Water has pato: acidity and pato: salinity; Drought and Heat Wave have pato: duration; Precipitation has pato: intensity; Ice has pato: radiation reflection quality (albedo). Natural and anthropogenic processes can change such qualities (attributes). For example, Human Activity can increase the Concentration of Carbon Dioxide in the Atmosphere, the Extent of Permafrost, and pH of the Sea Water.

Qualities inhered in material entities are measured by devices, and their values are commonly reported and analyzed by meteorologists and climatologists. The value and units of these variables can be modeled with the cco: Information Content Entity class which subsumes several classes for data and information modeling described above. Weather and climate data modeled with these classes can readily be integrated, facilitating their transfer and reuse. The 'abnormal' aspects of the Climate System such as Ocean Heat Content Anomaly, Specific Humidity Anomaly, Medieval Climate Anomaly, Cool Little Ice Age Anomaly, Land Surface Temperature Anomaly, and Ocean Surface Temperature Anomaly are modeled using the pato: abnormal class; a subclass of the pato: deviation (from_normal) class. Entities are related to their qualities through the 'ro: *has quality* object property.

Similar to material entities, occurrents (e.g., processes) such as Precipitation, Flood, Extreme Climate Event, and Monsoon have qualities such as pato: rate, pato: duration, pato: recurrent, and pato: frequency which are modeled under the pato: process quality class. BFO also provides the process profile class (a process) which proved handy for modeling the change in the rate of occurrence of adverse extreme climate events (e.g., Drought and Heat Wave) over time and the rate profile of Melting of Glaciers over a period of time.

Material-independent continuants also are characterized with realizable entities such as disposition and role. The bfo: disposition is inherent in material entities because of their physical make-up, such as composition, structure, and texture. For example, Disaster and Risk are dispositions of an Extreme Climate Event. Health, Disease, Right, Security, Well-being, and Vulnerability to Climate-related Extreme Event are human dispositions. These dispositions realize through certain processes. For example, Adverse Climate Event bfo: *realizes* Vulnerability to Heat Wave and Vulnerability to Flood. The cco: Electromagnetic Radiation Property class (a subclass of bfo: disposition) allowed defining CSO dispositions such as the Opacity of Soil, Transparent Ice, Radiation Absorptivity of Greenhouse Gas, Emissivity of Greenhouse Gas, Surface Albedo (under cco: Radiation Reflectivity), and Cosmic Ray Shielding Disposition of Ozone.

Each climate system component may also have a function which, as a subclass of disposition, is also a bfo: realizable entity. The CSO functions include Ecosystem Function, Generating Energy Function (of the Sun), and Regulating Earth Climate Function (of Ocean). The process of Nuclear Fusion in the Sun bfo: *realizes* the Generating Energy Function in the Sun. CCO provides a large number of classes for defining functions for artifacts, such as Sensor Artifact Function and Measurement Artifact Function. The bfo: realizable entity also includes role for continuant entities. These allow defining different kinds of bfo: role in CSO such as Policy Making Role, Carbon Storage Role, Driver of Climate Change Role, Driver of Deep Ocean Water Circulation Role, Carbon Sink Role, Proxy Role, and Greenhouse Gas Role. Processes realize roles. For example, Emission of Greenhouse Gas bfo: *realizes* Greenhouse Gas Role. Volcanic Eruption bfo: *realizes* the Aerosol Role for Volcanic Ash. Sampling from Ice Core bfo: *realizes* the Temperature Proxy Role for the Ice Core (a subclass of Sample). Nuclear fusion in the Sun bfo: *realizes* the 'Driver of Climate Change Role' for the Sun. Other processes realize dispositions. For example, Developing and Deploying Technology or Maintaining Stable Energy Supply bfo: *realizes* Energy Security. Some processes realize specific functions. Soil Moisture Drought cco: *has output* Reduced Ecosystem Function. Material entities relate to their realizable entities through the ro: *has disposition*, ro: *has function*, and ro: *has role* object properties.

The material and immaterial (i.e., independent continuant) parts of the climate system continuously interact through processes over time. The bfo: process provides mechanisms for the bfo: independent continuant system parts to interact. CCO subsumes the bfo: process by defining cco: Act, cco: Change, cco: Effect, cco: Natural Process, and cco: Stasis classes which we have used to define various dynamic aspects of CSO domain ontology. CSO classifies all anthropogenic activities, such as Fossil Fuel Emission, Emission of Halocarbon, and Industrial Activity, under the Human Activity class which is an indirect subclass of cco: Act. A large number of intentional anthropogenic activities are defined in CSO by subsuming the cco: Act class. These include the Evaluating Policies, Disaster Risk Management, Reduction of Disaster Risk, and Maintaining Stable Energy Supply classes. **Figure 2** displays the interactions of selected human activities in the climate system that are modeled in CSO.

The Climate System Ontology defines many dynamic processes that bring change in the components of the climate system. These changes, modeled as subclasses of the cco: Change class, include Change in Humidity and Change in Precipitation in the atmosphere, Varying Ice Area and Varying Snow Area in the Cryosphere, Change in the Storage of Groundwater, Change in Ocean Water Salinity, Change in Sea level in the hydrosphere, Land Cover Change, and Change in Surface Roughness in the lithosphere. The cco: Effect is used in CSO to define the Adverse Environmental Effect (and its

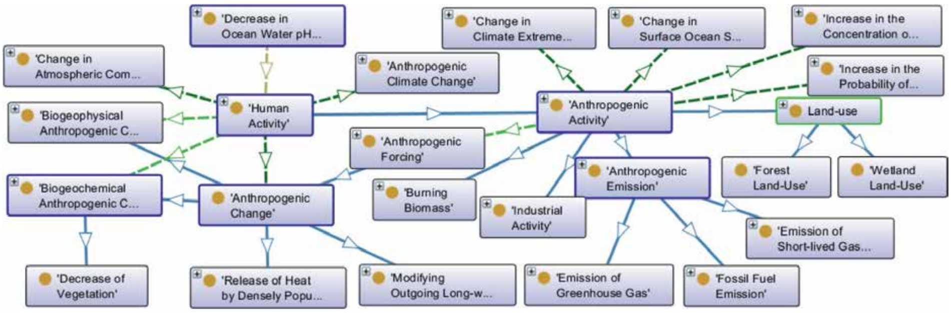

'Decrease in Ocean Water pH...
'Change in Climate Extreme...
'Change in Surface Ocean S...
'Increase in the Concentration o...
'Change in Atmospheric Com...
'Anthropogenic Climate Change'
'Increase in the Probability of...
'Biogeophysical Anthropogenic C...
'Human Activity'
'Anthropogenic Activity'
Land-use
'Anthropogenic Forcing'
'Biogeochemical Anthropogenic C...
'Anthropogenic Change'
'Burning Biomass'
'Anthropogenic Emission'
'Forest Land-Use'
'Wetland Land-Use'
'Industrial Activity'
'Emission of Short-lived Gas...
'Decrease of Vegetation'
'Release of Heat by Densely Popu...
'Modifying Outgoing Long-w...
'Emission of Greenhouse Gas'
'Fossil Fuel Emission'

Figure 2.
A model of human activities. Human Activity is cause of Change in Atmospheric Composition, process starts Biogeophysical Anthropogenic Change (with Decrease in Evapotranspiration and Decrease in Land Surface Net Radiation subclasses), process starts Biogeochemical Anthropogenic Change (with the Decrease of Vegetation and Decrease in Soil Carbon Stocks subclasses), and is cause of Anthropogenic Change (a subclass of Anthropogenic Forcing). Anthropogenic Activity is-a Human Activity. Anthropogenic Activity is cause of Change in Climate Extremes, Change in Surface Ocean Salinity, Increase in the Concentration of Greenhouse Gases, and Increase in the Probability of Occurrence of Heat Waves. The subclasses of Anthropogenic Activity include Industrial Activity, Anthropogenic Emission, and Land-use. Solid arrows point to subclasses.

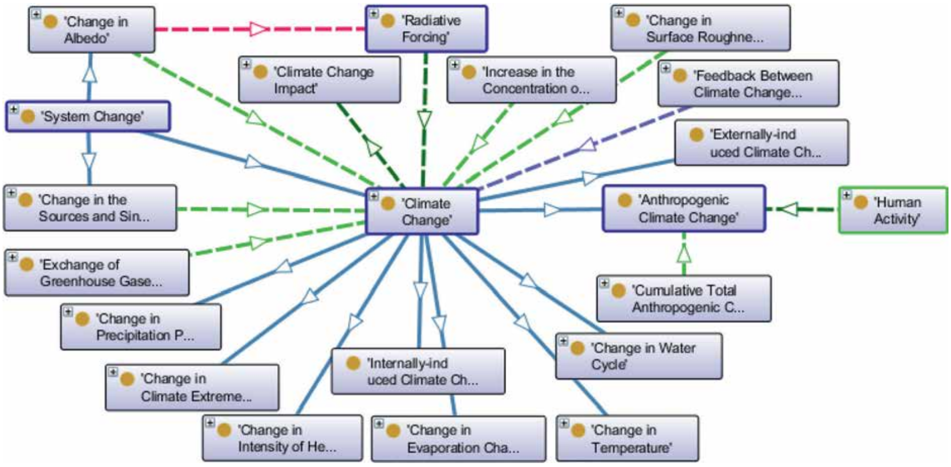

'Change in Albedo'
'Radiative Forcing'
'Change in Surface Roughne...
'Climate Change Impact'
'Increase in the Concentration o...
'Feedback Between Climate Change...
'System Change'
'Externally-induced Climate Ch...
'Change in the Sources and Sin...
'Climate Change'
'Anthropogenic Climate Change'
'Human Activity'
'Exchange of Greenhouse Gase...
'Cumulative Total Anthropogenic C...
'Change in Precipitation P...
'Change in Water Cycle'
'Internally-induced Climate Ch...
'Change in Climate Extreme...
'Change in Intensity of He...
'Change in Evaporation Cha...
'Change in Temperature'

Figure 3.
A model of climate change and its relation to other changes. Change in Albedo positively regulates (i.e., increases the frequency, magnitude, and rate of) Radiative Forcing and process starts Climate Change. Radiative Forcing is cause of Climate Change. Change in Albedo, Climate Change, and Change in the Sources and Sinks of Carbon are subclasses of System Change. Change in the Sources and Sinks of Carbon process starts Climate Change. Exchange of Greenhouse Gases Between Land and Atmosphere process starts Climate Change. Increase in the Concentration of Greenhouse Gases and Change in Surface Roughness process starts Climate Change. Feedback Between Climate Change and Atmospheric Concentration of Trace Gas has process part some Climate Change. Subclasses (subtypes) of Climate Change include Change in Precipitation Pattern, Change in Climate Extremes, Change in Intensity of Heavy Precipitation Over Land Regions, Internally-induced Climate Change, Change in Evaporation Characteristics, Change in Temperature, Change in Water Cycle, and Anthropogenic Climate Change. Cumulative Total Anthropogenic CO2 Emission process starts Anthropogenic Climate Change. Human Activity is cause of Anthropogenic Climate Change.

many subclasses) and Climate Change Impact classes and its subclasses (e.g., Impact on Infrastructure, Impact on Species, and Impact on Cultural Assets). **Figure 3** shows the inter-relationships among a few system changes, including the climate change.

Interactions among the components of the climate system are modeled under the cco: Natural Process class. These include Cloud Formation, Precipitation, and Wind in

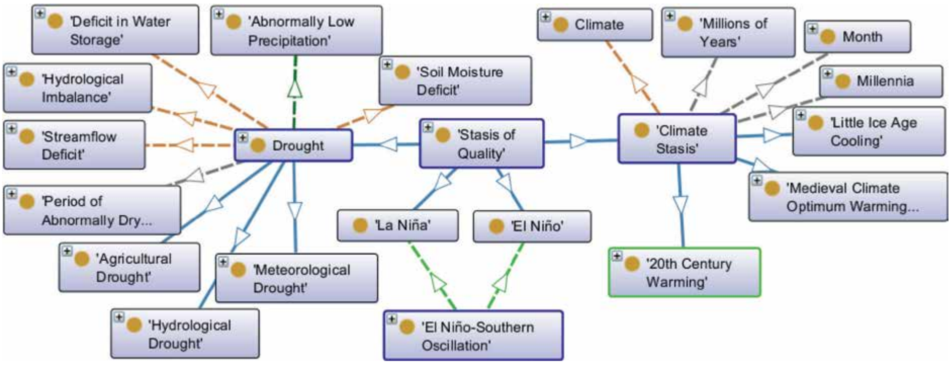


Figure 4.
Representing Drought as a cco: Stasis. Drought is-a cco: Stasis of quality. Drought has output Deficit in Water Storage, has output Hydrological Imbalance, has output Soil Moisture Deficit, has output Streamflow Deficit, is cause of Abnormally Low Precipitation, and occurs on Period of Abnormally Dry Weather. Agricultural Drought, Hydrological Drought, and Meteorological Drought are sub-types of Drought. Climate Stasis is-a cco: Stasis of quality. Climate Stasis has output Climate; occurs on Millions of Years or Month, or Millenia. Subclasses of Climate Stasis include the 20th Century Warming, Medieval Climate Optimum Warming, and Little Ice Age Cooling.

the atmosphere, Accumulation of Snow, Freezing, and Retreat of Glaciers in the cryosphere, Evapotranspiration and Photosynthesis in the biosphere, and Land Carbon Uptake from Atmosphere and Interception of Infiltrated Water by Vegetation classes in the lithosphere. Some climate system concepts such as 20th Century Warming, Little Ice Age Cooling, Drought, El Niño, and La Niña are classified under the cco: Stasis of Quality class. These are shown in **Figure 4**.

5. Discussion

In the previous section, we presented some modeling artifacts of our Climate System Ontology and demonstrated how classes and object properties of upper ontologies enabled formal modeling of some aspects of the climate system. In this section, we discuss the intricacies of our modeling of the complex climate system concepts such as solar radiation, feedback, climate change impact, enhanced greenhouse effect, hydrological cycle, oscillation, and radiative forcing. We also expand our modeling to include complex interactions among the climate system's components (e.g., Atmosphere–Hydrosphere, Land-Ice, Ice-Ocean, and Soil-Biosphere).

Internal and external forcings continuously evolve the climate system over a wide range of temporal and spatial scales. Major perturbations in the radiative balance lead to self-organization, by creating and changing climate patterns. Nonlinear feedback mechanisms continuously amplify or dampen processes to allow system components to adapt to new changes. The system reorganizes to maintain its identity, structure, and function through new processes and patterns (e.g., more frequent extreme events) or by building resilience. Below, we elaborate on the modeling of some of these complex interactions.

Figure 5 displays the transformation of the Solar Radiation as it continuously enters and exits the climate system. The Solar Radiation class is modeled as a subclass of cco: Electromagnetic Wave Process (a subclass of cco: Wave Process).

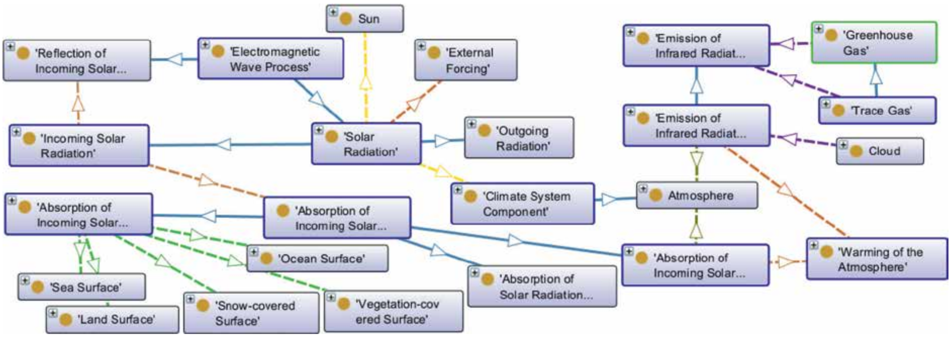

Sun
'Reflection of Incoming Solar...
'Electromagnetic Wave Process'
'External Forcing'
'Emission of Infrared Radiat...
'Greenhouse Gas'
'Trace Gas'
'Emission of Infrared Radiat...
Cloud
'Incoming Solar Radiation'
'Solar Radiation'
'Outgoing Radiation'
'Climate System Component'
Atmosphere
'Absorption of Incoming Solar...
'Absorption of Incoming Solar...
'Ocean Surface'
'Absorption of Solar Radiation...
'Absorption of Incoming Solar...
'Warming of the Atmosphere'
'Sea Surface'
'Land Surface'
'Snow-covered Surface'
'Vegetation-covered Surface'

Figure 5.
A model of different types of Radiation in CSO. Solar Radiation is-a cco: Electromagnetic wave process, cco: Has input (i.e., involves) Sun, cco: Has input Climate System Component, and cco: Process starts (i.e., causes) External Forcing. The Outgoing Radiation and Incoming Solar Radiation classes are subclasses of Solar Radiation. Reflection of Incoming Solar Radiation is-a cco: Electromagnetic wave process. Incoming Solar Radiation ro: Directly positively regulates (i.e., increases frequency, magnitude, and rate of) Reflection of Incoming Solar Radiation. Incoming Solar Radiation ro: Directly positively regulates Absorption of Incoming Solar Radiation. Absorption of Incoming Solar Radiation has subclass Absorption of Incoming Solar Radiation by the Atmosphere that ro: Directly positively regulates Warming of the Atmosphere and occurs in the Atmosphere. Absorption of Incoming Solar Radiation also has subclass Absorption of Incoming Solar Radiation by Surface which occurs at Sea Surface, Land Surface, Snow-covered Surface, and Vegetation-covered surface. Emission of Infrared Radiation has input Cloud and has subclass Emission of Infrared Radiation in All Directions with Greenhouse Gas and other trace gases as input. Emission of Infrared Radiation occurs in the Atmosphere and process starts (i.e., causes) Warming of the Atmosphere.

It subsumes the Incoming Solar Radiation and Outgoing Solar Radiation classes. The cco: Electromagnetic Wave Process also subsumes the Reflection of Incoming Solar Radiation, Absorption of Radiation, Return of the Radiation Absorbed by the Surface, Emission of Radiation, and Scattering of Part of the Incoming Solar Radiation classes.

The logical modeling artifacts provided by the imported ontologies also allowed CSO to efficiently model complex system features such as nonlinear dynamics through feedbacks. **Figure 6** shows an example of a positive feedback through which the water vapor (a major greenhouse gas), produced through the evaporation of ocean water, cycles through the atmosphere. The increased concentration of the water vapor leads to an increase in atmospheric temperature, which amplifies the original evaporation process, and leads to increased concentration of water vapor, bringing more heat in the atmosphere. The cyclical Feedback class and its subclasses are modeled in CSO as a subclass of the Cycle class, which is modeled as a subclass of the Fiat Process Part class (a subclass of cco: Change).

Various impacts brought by climate change are also modeled in CSO using several modeling artifacts of CCO. **Figure 7** shows some of the impacts of extreme climate events and represents how these events realize the vulnerability (a disposition) of communities to such events. Several subclasses of the Climate Change Impact class explicitly define specific types of impacts (not expanded in the diagram).

The concepts of natural and enhanced greenhouse effects are shown in the model of **Figure 8**. The figure shows the natural and anthropogenic processes and material entities that cause or are involved in these two types of greenhouse effects.

The hydrological cycle is a major global process in the climate system. It involves numerous processes that occur in different components of the climate

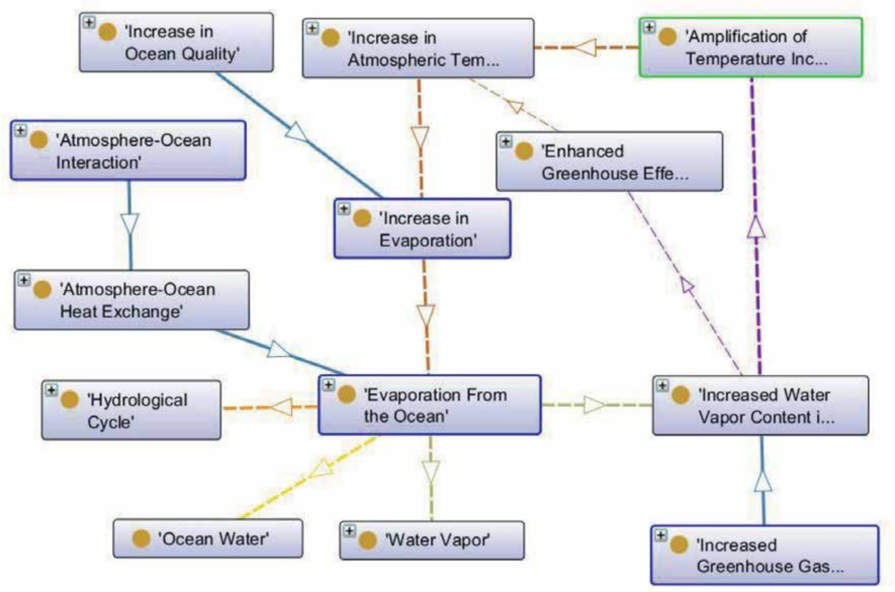

'Increase in Ocean Quality'
'Increase in Atmospheric Tem...
'Amplification of Temperature Inc...
'Atmosphere-Ocean Interaction'
'Enhanced Greenhouse Effe...
'Increase in Evaporation'
'Atmosphere-Ocean Heat Exchange'
'Hydrological Cycle'
'Evaporation From the Ocean'
'Increased Water Vapor Content i...
'Ocean Water'
'Water Vapor'
'Increased Greenhouse Gas...

Figure 6.
A model of a positive feedback in CSO. Increase in Atmospheric Temperature process starts (causes) Increase in Evaporation which process starts Evaporation from the Ocean. Evaporation From the Ocean has input Ocean Water and has output Water Vapor. Evaporation From the Ocean is part of process Hydrological Cycle. Evaporation from the ocean has output Increased Water Vapor Content in the Atmosphere which is an Increased Greenhouse Gas Concentration. Increased Water Vapor Content in the atmosphere 'is input of' Amplification of Temperature Increase and is input of Enhanced Greenhouse Effect. Each of these processes process starts Increase in Atmospheric Temperature which process starts Increase in Evaporation which restarts the cycle by amplifying the Evaporation from the Ocean.

system. The cyclical process starts by evaporation from bodies of water such as oceans, movement of the output water vapor through atmospheric circulations, condensation of the water vapor, cloud formation, and precipitation as rain or snow. Precipitation is followed by the interception of rain and snow by plants, infiltration of rain and melted snow into soil, soil evaporation, recharge of aquifers, surface runoff, and entry of streams back into oceans. **Figure 9** is a model of part of the hydrological cycle in CSO. The Hydrological Cycle class is modeled in CSO as a subclass of the Water Cycle, under the Fiat Process Part class (a subclass of cco: Change).

The three major subclasses of cco: Information Content Entity (see above) that represent data and information, in combination with the bfo: specifically dependent continuant that defines the system variables, and object properties such as ro: *concretize* that relates a variable (quality) to data and information and enable complete modeling of climate system data. Data modeled through these upper-level constructs enable integration of climate data and information. For example, numeric, graphic (map, plot), and textual (report) data related to specific occurrences (instances) of an El Niño or La Niña event, shown in the CSO model in **Figure 10** can be modeled with the cco: Information Content Entity.

Radiative forcing as a change in the incoming and outgoing radiative flux may be caused by changes in the concentration of anthropogenic greenhouse gases in the atmosphere because of human activities or solar cycles. CSO models Radiative

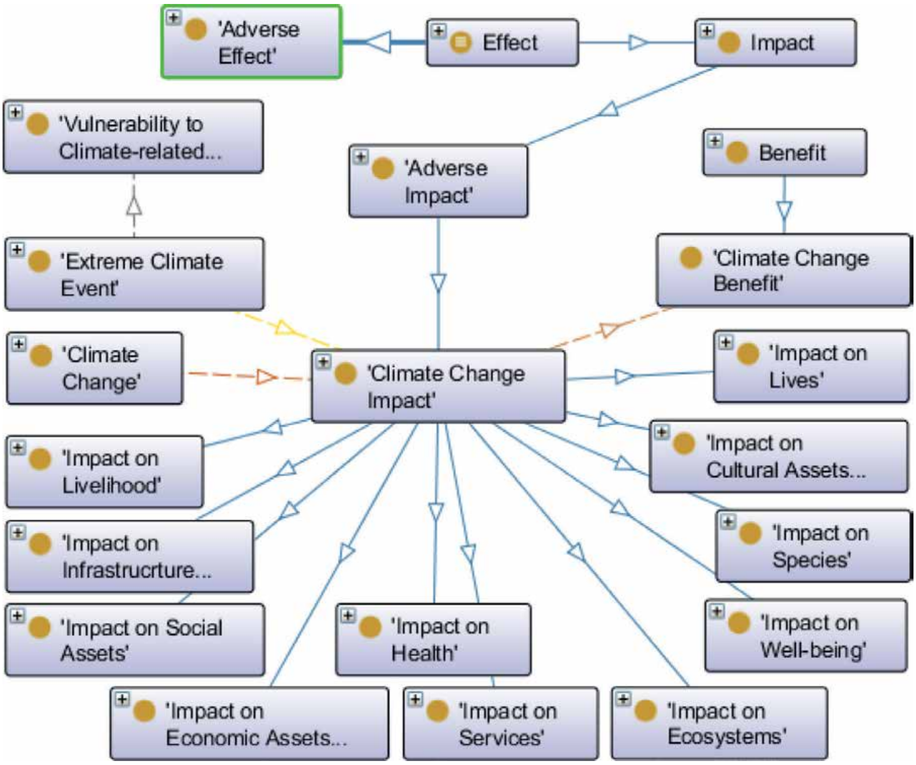


Figure 7.
A model of Climate Change Impact in CSO as an Adverse Impact, a subclass of Impact. The Impact class is-a subclass of Effect which is-a Adverse Effect (under cco: Effect). Different kinds of impacts are also shown as subclasses of the Climate Change Impact. Climate Change is cause of some Climate Change Impact. The Extreme Climate Event process starts (i.e., causes) Climate Change Impact and realizes Vulnerability to Climate-related Extreme Event.

Forcing and its underlying Positive Radiative Forcing using several cco object properties. **Figure 11** shows anthropogenic activities lead to anthropogenic forcing, causing increase in the global mean surface temperature which causes global mean surface warming. It also shows anthropogenic activities (e.g., emission of halocarbons) and increase in atmospheric opacity cause positive radiative forcing. Positive radiative forcing with input from greenhouse gases starts the process of enhanced greenhouse effect which leads to an increase in atmospheric temperature and ultimately the warming of atmosphere.

6. Summary

The Climate System Ontology (CSO) specifies the characteristics of different elements of the climate system and models the dynamic processes that affect the structure, behavior, and pattern of these elements at the micro- (component) and macro- (whole system) levels. For each process, the ontology identifies the components that change their attributes as they participate in the process as input and other processes that are caused by the process and produce their own output. Many of these processes are nonlinear, producing outputs that affect the original process that

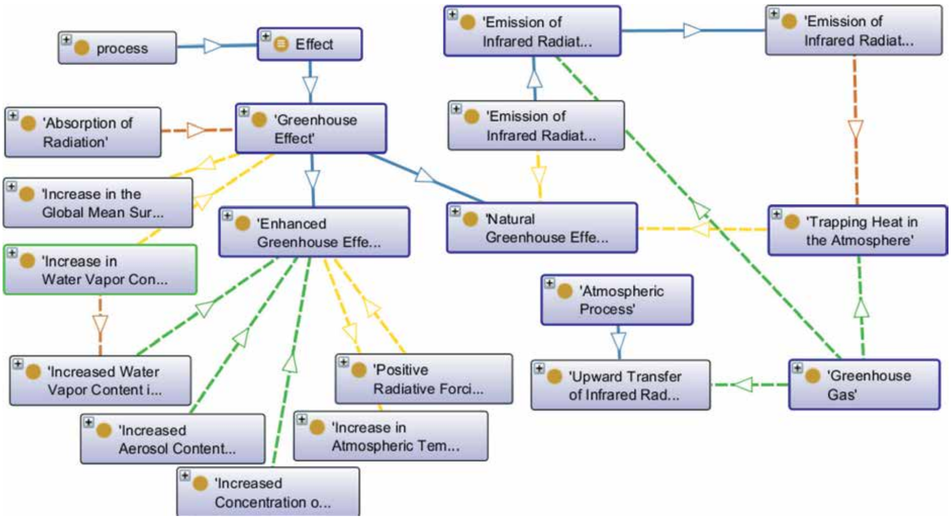

process
Effect
'Emission of Infrared Radiat...
'Emission of Infrared Radiat...
'Absorption of Radiation'
'Greenhouse Effect'
'Emission of Infrared Radiat...
'Increase in the Global Mean Sur...
'Enhanced Greenhouse Effe...
'Natural Greenhouse Effe...
'Trapping Heat in the Atmosphere'
'Increase in Water Vapor Con...
'Atmospheric Process'
'Increased Water Vapor Content i...
'Positive Radiative Forci...
'Upward Transfer of Infrared Rad...
'Greenhouse Gas'
'Increased Aerosol Content...
'Increase in Atmospheric Tem...
'Increased Concentration o...

Figure 8.

A model of Greenhouse Effect as a cco: Effect, under bfo: Process. Greenhouse Effect process starts (i.e., is cause of) some Increase in the Global Mean Surface Temperature. As a subclass of Greenhouse Effect, the Enhanced Greenhouse Effect process starts some Increase in Atmospheric temperature. Increased Aerosol Content, Increased Concentration of Greenhouse Gases, and Increased Water Vapor Content in the Atmosphere are input of Enhanced Greenhouse Effect. Increase in Water Vapor Content in the Atmosphere has output some Increased Water Vapor Content in the Atmosphere, and process starts some Greenhouse Effect. Emission of Infrared Radiation is cause of Trapping Heat in the Atmosphere. Greenhouse Gas is input of Trapping Heat in the Atmosphere and is input of Emission of Infrared Radiation. Positive Radiative Forcing process starts some enhanced Greenhouse Effect. Absorption of Radiation is cause of some Greenhouse Effect. Greenhouse gas is input of Upward Transfer of Infrared Radiation From Earth Surface to Higher Altitudes (an Atmospheric Process).

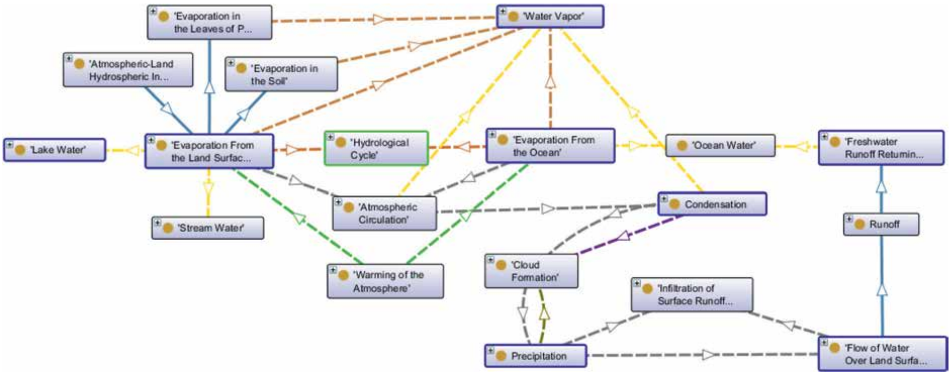

'Evaporation in the Leaves of P...
'Water Vapor'
'Atmospheric-Land Hydrospheric In...
'Evaporation in the Soil'
'Lake Water'
'Evaporation From the Land Surfac...
'Hydrological Cycle'
'Evaporation From the Ocean'
'Ocean Water'
'Freshwater Runoff Returnin...
'Atmospheric Circulation'
Condensation
'Stream Water'
Runoff
'Warming of the Atmosphere'
'Cloud Formation'
'Infiltration of Surface Runoff...
Precipitation
'Flow of Water Over Land Surfa...

Figure 9.

A model of Hydrological Cycle. Evaporation From the Ocean has input Ocean Water and has output Water Vapor. Evaporation From the Land Surface has input Lake Water and has input Stream Water. Evaporation from the Land Surface or Evaporation from the Ocean process starts (i.e., causes) Atmospheric Circulation. Evaporation From the Land Surface is-a Atmospheric–Land Hydrospheric Interaction. Evaporation in the Soil or Evaporation in the Leaves of Plants is subclasses of Evaporation From the Land Surface. Warming of the Atmosphere directly positively regulates (i.e., intensifies) Evaporation. Atmospheric Circulation process starts Condensation. Condensation has input Water Vapor and has output Cloud Droplet. Condensation process precedes, and process starts Cloud Formation. Cloud Formation process precedes, and process starts Precipitation. Precipitation process starts Flow of Water Over Land Surface or Through the Subsurface and Infiltration of Surface Runoff into Soil and Rock. Runoff, as a subclass of the Flow of Water Over Land Surface or Through the Subsurface, has subclass Freshwater Runoff Returning to the Oceans and has input (i.e., involves) Ocean Water. Ocean Water re-enters the hydrological cycle through the Evaporation From the Ocean. The cycle repeats the above processes.

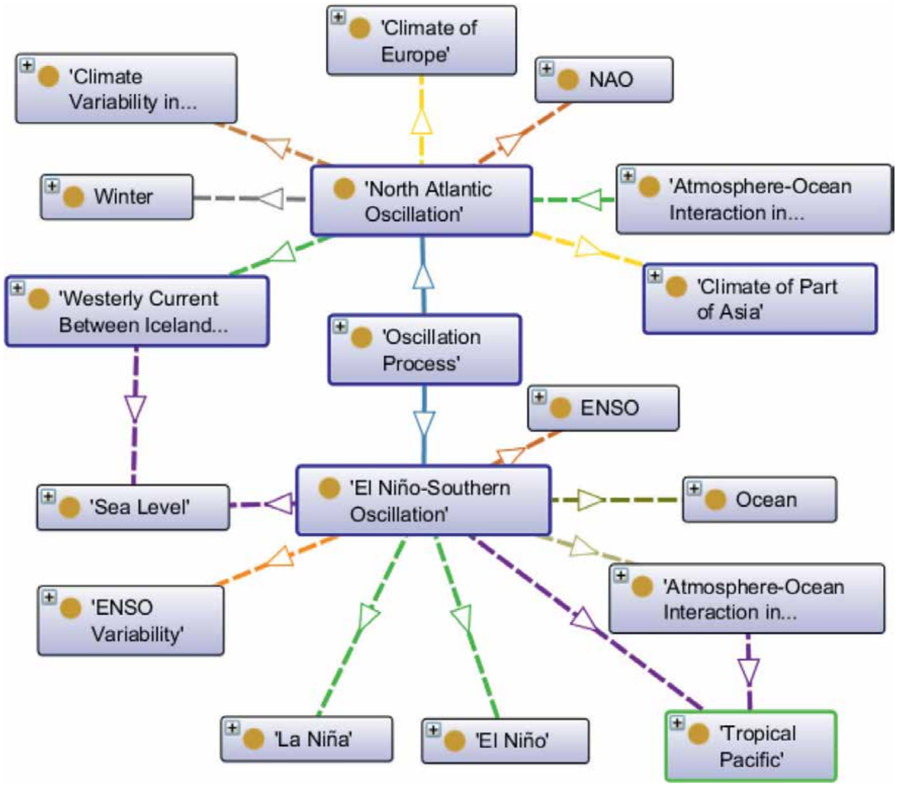

'Climate of Europe'
'Climate Variability in...
NAO
Winter
'North Atlantic Oscillation'
'Atmosphere-Ocean Interaction in...
'Westerly Current Between Iceland...
'Climate of Part of Asia'
'Oscillation Process'
ENSO
'Sea Level'
'El Niño-Southern Oscillation'
Ocean
'ENSO Variability'
'Atmosphere-Ocean Interaction in...
'La Niña'
'El Niño'
'Tropical Pacific'

Figure 10.
A model of oceanic oscillations. Two oscillations are shown. North Atlantic Oscillation designated by NAO realizes Climate of Europe; has output Climate Variability in Europe in Winter; occurs on Winter; process starts Westerly Current Between Icelandic Low Pressure and Azores High Pressure Areas (which occurs at Sea Level); realizes Climate of Part of Asia. Atmospheric–Ocean Interactions in North Atlantic process starts North Atlantic Oscillation. The El Niño-Southern Oscillation designated by ENSO; occurs in Ocean; occurs at Sea Level; occurs at Tropical Pacific; has quality ENSO Variability; process starts El Niño and process starts La Niña; caused by Atmosphere–Ocean Interaction in the Tropical Pacific that occurs at Tropical Pacific.

led to the output. The CSO models the dynamic interactions among system's major components as the solar radiative energy cycles through the system through various reflective, absorptive, and emissive processes. We described the climate system from a complex system perspective and modeled several of its dynamic processes based on this view.

We developed the Climate System Ontology by extending the class hierarchy and logic of a set of well-designed, top- and mid-level ontologies. The terminology used for the class definitions and relations used to model the Climate System Ontology are based on the IPCC and other sources of climate system knowledge. The use of the foundational logics of the imported upper-level ontologies in the development of the Climate System Ontology ensures interoperability with other ontologies that extend the same upper-level ontologies. We gave full descriptions of these upper-level ontologies and specified best practices for using them to build domain or application ontologies. We demonstrated, by providing several examples, how complex features in the climate system can be modeled in the Protégé editor. The ontology is publicly available in the GitHub cloud repository for extension by climate scientists to build their own application ontologies. The

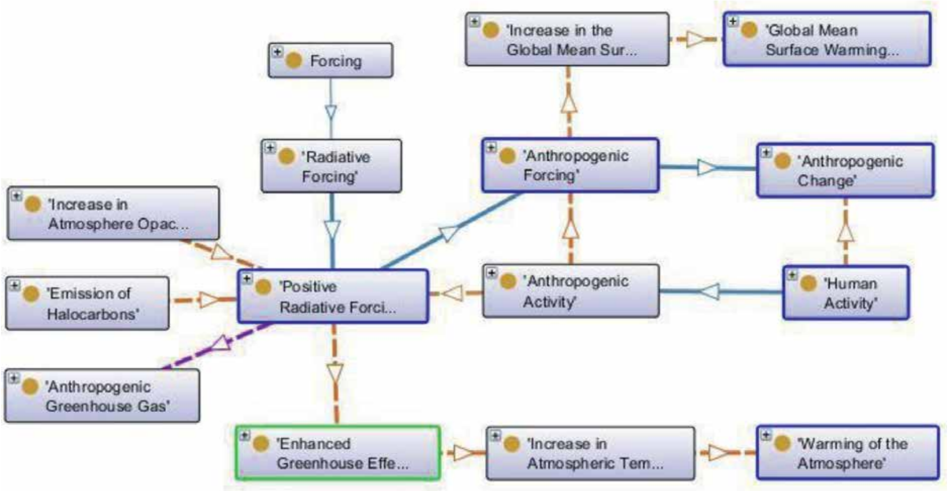

Forcing
'Increase in the Global Mean Sur...
'Global Mean Surface Warming...
'Radiative Forcing'
'Anthropogenic Forcing'
'Anthropogenic Change'
'Increase in Atmosphere Opac...
'Emission of Halocarbons'
'Positive Radiative Forci...
'Anthropogenic Activity'
'Human Activity'
'Anthropogenic Greenhouse Gas'
'Enhanced Greenhouse Effe...
'Increase in Atmospheric Tem...
'Warming of the Atmosphere'

Figure 11.
A model of part of the Positive Radiative Forcing process in CSO. Forcing is modeled as a subclass of cco: Change (not shown). Radiative Forcing is-a forcing and Positive Radiative Forcing is-a Radiative Forcing. Positive Radiative Forcing has input Anthropogenic Greenhouse Gas. Emission of Halocarbons and Increase in Atmospheric Opacity process starts Positive Radiative Forcing. Anthropogenic Forcing is-a Positive Radiative Forcing. Anthropogenic Forcing is cause of Increase in the Global Mean Surface Temperature which is cause of Global Mean Surface Warming. Human Activity is cause of Anthropogenic Change. Anthropogenic Activity is cause of Positive Radiative Forcing which process starts Enhanced Greenhouse Effect, which in turn causes Increase in Atmospheric Temperature, which is cause of Warming of the Atmosphere.

Climate System Ontology can be queried by decision and policymakers to discover the effects of different kinds of natural and anthropogenic processes that occur in the complex climate system.

Author details

Armita Davarpanah[1]*, Hassan A. Babaie[2] and Guanyu Huang[1]

1 Environmental and Health Sciences, Spelman College, Atlanta, GA, USA

2 Department of Geosciences, Georgia State University, Atlanta, GA, USA

*Address all correspondence to: adavarpa@spelman.edu

References

[1] Ahrens CD. Meteorology Today: An Introduction to Weather, Climate, and the Environment. Thomson/Brooks/Cole: Belmont, CA; 2007. p. 688

[2] Holland JH. Emergence. Philosophica. 1997;**59**(1):11-40

[3] Munoz YJ, de Castro LN. Self-organization and emergence in artificial life: Concepts and illustrations. Journal of Experimental & Theoretical Artificial Intelligence. 2009;**21**(4):273-292

[4] Ehrlich P, Raven P. Butterflies and plants: A study in coevolution. Evolution. 1964;**18**(4):586-608

[5] AR5. Synthesis Report: Climate Change 2014. The Intergovernmental Panel on Climate Change. 2014. Available online: https://www.ipcc.ch/report/ar5/syr/ [Accessed: March 28, 2022].

[6] Gruber N, Bakker DCE, DeVries T, Gregor L, Hauck J, Landschützer P, et al. Trends and variability in the ocean carbon sink. Nature Reviews & Earth Environment. 2023;**4**:119-134. DOI: 10.1038/s43017-022-00381-x

[7] Collins M, Knutti R, Arblaster J, Dufresne J-L, Fichefet T, Friedlingstein P, et al. Long-term climate change: Projections, commitments and irreversibility. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, et al, editors. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press; 2013

[8] Lüthi D, Le Floch M, Bereiter B, Blunier T, Barnola J-M, Siegenthaler U, et al. High-resolution carbon dioxide concentration record 650,000-800,000 years before present. Nature. 2008;**453**:379-382. DOI: 10.1038/nature06949

[9] Feldman DR, Collins WD, Gero PJ, Torn MS, Mlawer EJ, Shippert TR. Observational determination of surface radiative forcing by CO_2 from 2000 to 2010. Nature. 2015;**519**:339-343. DOI: 10.1038/nature14240

[10] Rial JA, Pielke RA, Beniston M, Claussen M, Canadell J, Cox P, et al. Nonlinearities, feedbacks and critical thresholds within the Earth's climate system. Climatic Change. 2004;**65**:11-38. DOI: 10.1023/B:CLIM.0000037493.894 89.3f

[11] Climate Change. Climate change evidence & causes - update 2020. An overview from the Royal Society and the US National Academy of Sciences. 2020. Available online: https://royalsociety.org/~/media/royal_society_content/policy/projects/climate-evidence-causes/climate-change-evidence-causes.pdf [Accessed: March 28, 2022]

[12] Gilbert L. Concepts and Applications of Climatology. New York, NY: Syrawood Publishing House; 2019. p. 216

[13] Somerville RCJ, Hassol SJ. Communicating the science of climate change. Physics Today. 2011;**64**(10):48-53

[14] Sullivan D. Climatology. New York, NY: Callisto Reference; 2019. p. 219

[15] AR6. Sixth Assessment Report. Climate Change 2022: Impacts, Adaptation and Vulnerability Six Assessment Report. The working Group II contribution to the Six Assessment

Report. 2022. Available online: www.ipcc.ch/assessment-report/ar6/ [Accessed: March 28, 2022]

[16] OWL. The W3C Web Ontology Language (OWL). 2004. Available online: www.w3.org/OWL/ [Accessed: March 10, 2022]

[17] Baader F. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The Description Logic Handbook – Theory, Implementation, and Applications. New York: Cambridge University Press; 2007. p. 602

[18] Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology. Cambridge MA, USA: MIT Press; 2015. p. 248

[19] BFO. Basic Formal Ontology. 2022. Available online: https://basic-formal-ontology.org/ [Accessed: June 28, 2022]

[20] CCO. Available online: https://github.com/CommonCoreOntology/CommonCoreOntologies)

[21] Rudnicki R. Modeling Information with the Common Core Ontologies. Buffalo, NY: CUBRC Inc.; 2017. Available online: https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-cubrc_inc_003.pdf. Accessed March 28, 2022

[22] Rudnicki R. An Overview of the Common Core Ontologies. Buffalo, NY: CUBRC Inc.; 2019. Available online: https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-cubrc_inc_004.pdf. [Accessed March 28, 2022]

[23] RO. Relation Ontology (RO). 2008. Available online: https://github.com/oborel/obo-relations [Accessed: June 15, 2022]

[24] PATO. The 'Phenotype and Trait Ontology'. 2022. Available online: https://raw.githubusercontent.com/pato-ontology/pato/master/pato.owl

[25] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. 1993;**5**(2):199-220

[26] Baede APM, Ahlonsou E, Ding Y, Schimel D, Bolin B, Pollonais S. The climate system: An overview. In: Houghton JT, Ding Y, Grigss DJ, Noguer M, Linden PJ, Van Der D, et al, editors. Climate Change 2001: The Scientific Basis. Cambridge: Cambridge University Press; 2001

[27] Trenberth KE, Fasullo JT, Kiehl JT. Earth's global energy budget. Bull. Amer. Meteor. Soc. 2009;**90**:311-323

[28] National Academy of Sciences. Climate Change and Ecosystems. Washington, DC: The National Academies Press; 2019. DOI: 10.17226/25504

[29] Fröhlich C, Lean JL. The Sun's total irradiance: Cycles, trends and related climate change uncertainties since 1976. Geophysical Research Letters. 1998;**25**:4377-4380

[30] Lean JL, Rind D. Climate forcing by changing solar radiation. Journal of Climate. 1998;**11**:3069-3094

[31] Kiehl JT, Trenberth KE. Earth's annual global mean energy budget. Bull. Am. Met. Soc. 1997;**78**:197-208

[32] Cassia R, Nocioni M, Correa-Aragunde N, Lamattina L. Climate change and the impact of greenhouse gasses: CO_2 and NO, friends and foes of plant oxidative stress. Frontiers in Plant Science. 2018;**9**:273. DOI: 10.3389/fpls.2018.00273

[33] Boucher O, Randall D, Artaxo P, Bretherton C, Feingold G, Forster P,

et al. Clouds and aerosols. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, et al, editors. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press; 2013

[34] Gettelman A, Forster PM d F, Fujiwara M, Fu Q, Vomel H, Gohar LK, et al. Radiation balance of the tropical tropopause layer. Journal of Geophysical Research. 2004;**109**:D07103. DOI: 10.1029/2003JD004190

[35] Pierrehumbert RT. Infrared radiation and planetary temperature. Physics Today. 2011;**64**:33-38

[36] Rypdal K. Global temperature response to radiative forcing: Solar cycle versus volcanic eruptions. Journal of Geophysical Research. 2012;**117**:D06115. DOI: 10.1029/2011JD017283

[37] Stocker TF, Clarke GKC, Le Treut H, Lindzen RS, Meleshko, VP, Mugara RK, et al. Physical climate processes and feedbacks. In: Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, editors. IPCC, 2001: Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. New York, NY: Cambridge University Press; 2001. pp. 417-470

[38] Raghuraman SP, Paynter D, Ramaswamy V. Anthropogenic forcing and response yield observed positive trend in Earth's energy imbalance. Nature Communications. 2021;**12**:4577. DOI: 10.1038/s41467-021-24544-4

[39] Hall-Spencer JM, Rodolfo-Metalpa, Martin RS, Ransome E, Fine M, Turner SM, et al. Volcanic carbon dioxide vents show ecosystem effects of ocean acidification. Nature. 2008;**454**:96-99

[40] Riebesell U. Climate change: Acid test for marine biodiversity. Nature. 2008;**454**:46-47

[41] Wuebbles DJ, Easterling DR, Hayhoe K, Knutson T, Kopp RE, Kossin JP, et al. Our globally changing climate. In: Wuebbles DJ, Fahey DW, Hibbard KA, Dokken DJ, Stewart BC, Maycock TK, editors. Climate Science Special Report: Fourth National Climate Assessment. Vol. I. Washington, DC, USA: U.S. Global Change Research Program; 2017. pp. 35-72. DOI: 10.7930/J08S4N35

[42] Jain PC. Greenhouse effect and climate change: Scientific basis and overview. Renewable Energy. 1993;**3**(4-5):403-420

[43] Kheshgi HS, White BS. Does recent global warming suggest an enhanced greenhouse effect? Climatic Change. 1993;**23**:121-139. DOI: 10.1007/BF01097333

[44] Feldl N, Roe GH. The nonlinear and nonlocal nature of climate feedbacks. Journal of Climate. 2013;**26**:8289-8304

[45] Meehl GA, Washington WM, Arblaster JM, Hu A, Teng H, Tebaldi C, et al. Climate system response to external forcings and climate change projections in CCSM4. Journal of Climate. 2012;**25**(11):3661-3683. DOI: 10.1175/jcli-d-11-00240.1

[46] Wang C, Deser C, Yu J-Y, DiNezio P, Clement A. El Niño-southern oscillation (ENSO): A review. In: Glymn P, Manzello D, Enochs I, editors. Coral Reefs of the Eastern Pacific. Dordrecht: Springer Science Publisher; 2016. pp. 85-106

[47] Dong L, McPhaden MJ. The role of external forcing and internal variability in regulating global mean surface temperatures on decadal timescales. Environmental Research Letters. 2017;**12**:034011

[48] Luber G, McGeehin M. Climate change and extreme heat events. American Journal of Preventive Medicine. 2008;**35**(5):429-435

[49] Myers KF, Doran PT, Cook J, Kotcher JE, Myers TA. Consensus revisited: Quantifying scientific agreement on climate change and climate expertise among earth scientists 10 years later. Environmental Research Letters. 2021;**16**(10):104030. DOI: 10.1088/1748-9326/ac2774

[50] Benbya H, Nan N, Tanriverdi H, Yoo Y. Complexity and information systems research in the emerging digital world. MIS Quarterly. 2020;**44**(1):1-17. Special Issue: Complexity & IS Research

[51] Holland JH. Hidden Order: How Adaptation Builds Complexity. Reading, MA: Addison-Wesley; 1995

[52] Glansdorff P, Prigogine I. Thermodynamic Study of Structure, Stability and Fluctuations. New York: Wiley; 1978

[53] Bak P, Tang C, Wiesenfeld K. Self-organized criticality. Physical Review A. 1988a;**38**:364

[54] Crucifix M. Oscillators and relaxation phenomena in Pleistocene climate theory. Philosophical Transactions of the Royal Society A. 2012;**370**:1140-1165. DOI: 10.1098/rsta.2011.0315

[55] Roli A, Villani M, Filisetti A, et al. Dynamical criticality: Overview and open questions. Journal of Systems Science and Complexity. 2018;**31**:647-663. DOI: 10.1007/s11424-017-6117-5

[56] Kauffman SA. The Origins of Order: Self-Organization and Selection in Evolution. New York: Oxford University Press; 1993

[57] Lewin R. Complexity: Life at the Edge of Chaos. Chicago: University of Chicago Press; 1992

[58] Johnson GC, Lyman JM. Warming trends increasingly dominate global ocean. Nature Climate Change. 2020;**10**:757-761. DOI: 10.1038/s41558-020-0822-0

[59] Watson AJ, Schuster U, Shutler JD, Holding T, Ashton IGC, Landschützer P, et al. Revised estimates of ocean-atmosphere CO_2 flux are consistent with ocean carbon inventory. Nature Communications. 2020;**11**:4422. DOI: 10.1038/s41467-020-18203-3.

[60] De Wolf T, Holvoet T. Emergence versus self-organization: different concepts but promising when combined. In: Brueckner SA, Di Marzo SG, Karageorgos A, Nagpal R, editors. Engineering Self-Organising Systems. ESOA 2004. Lecture Notes in Computer Science. Vol. 3464. Berlin, Heidelberg: Springer; 2005. DOI: 10.1007/11494676_1

[61] Heylighen F. The science of self-organisation and adaptivity. In: The Encyclopedia of Life Support Systems. Paris, France: UNESCO Publishing-Eolss Publishers; 2002

[62] Byeon JH. Non-equilibrium thermodynamics approach to the change in political systems. System Research and Behavioral Science. 1999;**16**:283-291

[63] Langton CG. Computation at the edge of chaos: Phase transitions and emergent computation. Physica D. 1990;**42**:12-37

[64] Bak P, Tang C, Wiesenfeld K. Scale invariant spatial and temporal fluctuations in complex systems. Random Fluctuations and Pattern Growth: Experiments and Models. 1988b;**157**:329-335. ISBN: 978-0-7923-0073-1

[65] Prigogine I, Stengers I. Order out of Chaos. New York: Bantam; 1984

[66] Prigogine I, Nicolis G, Babylontz A. Thermodynamics of evolution. Physics Today. 1971;**25**(11):23

[67] Gershenson C. Design and Control of Self-Organizing Systems. PhD Dissertation. Belgium: Vrije Universiteit Brussel; 2007 http://cogprints.org/5442/1/thesis.pdf

[68] Juarrero A. Dynamics in action: Intentional behavior as a complex system. Emergence. 2000;**3**(2):24-57

[69] Mitchel SD. Emergence: Logical, functional, and dynamical. Synthese. 2012;**185**:171-186

[70] Jacobson MJ, Kapur M, So J-J, Lee J. The ontologies of complexity and learning about complex systems. Instructional Science. 2011;**39**:763-883

[71] Couzin ID, Krause J. Self-organization and collective behavior in vertebrates. Advances in the Study of Behavior. 2003;**33**:1-75

[72] BFO Standard. ISO/IEC PRF 21838-2.2 Information Technology — Top-Level Ontologies (TLO) — Part 2: Basic Formal Ontology (BFO). 2021. https://www.iso.org/standard/74572.html [Accessed: June 10, 2022]

[73] BFO users. 2022. Available online: https://basic-formal-ontology.org/users.html [Accessed: June 27, 2022]

[74] Github. Open Source Repository for Software Development and Version Control. 2023. Available online: http://github.com

[75] Musen MA. The Protégé project: A look Back and a look forward. AI Matters. 2015;**1**:4-12. DOI: 10.1145/2757001.2757003

[76] Protégé. A Free, Open-Source Ontology Editor and Framework for Building Intelligent Systems. 2022. Available online: https://protege.stanford.edu/ [Accessed: December 10, 2021]

[77] Rudnicki R, Smith B, Malyuta T, Mandrick COLW. Best Practices of Ontology Development. White Paper. Buffalo, NY: CUBRC; 2016. Retrieved March 30, 2022 from: https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-cubrc_inc_002.pdf

[78] Partridge C, Mitchell A, Cook A, Sullivan J, West M. A Survey of Top-Level Ontologies - to Inform the Ontological Choices for a Foundation Data Model. 2020. DOI: 10.17863/CAM.58311. Available from: https://www.cdbb.cam.ac.uk/files/a_survey_of_top-level_ontologies_lowres.pdf. Constructioninnovationhub.org.uk, UK Research and Innovation [Accessed: June 24, 2022]

[79] RDFS. RDF Schema 1.1. W3C Recommendation 25 February 2014. 2014. https://www.w3.org/TR/rdf-schema/

[80] Motik B, Patel-Schneider PF, Parsia B, Bock C, Fokoue A, Haase P, et al. OWL 2 web ontology language: Structural specification and functional-style syntax. W3C recommendation. 2009;**27**:159

[81] Shearer R, Motik B, Horrocks I. HermiT: A highly-efficient OWL reasoner. In: Proceedings of the

5th OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, Oct 2 26-27. 2008

[82] Otte JN, Kiritsi D, Mohd Ali M, Yang R, Zhang B, Rudnicki R, et al. An ontological approach to representing the product life cycle. Applied Ontology. 2019;**14**(2):179-197

[83] Rodríguez-Fonseca B, Suárez-Moreno R, Ayarzagüena B, López-Parages J, Gómara I, Villamayor J, et al. A review of ENSO influence on the North Atlantic. A non-stationary signal. Atmosphere. 2016;**7**:87. DOI: 10.3390/atmos7070087

[84] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biology. 2005;**6**(5):R46

[85] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied Ontology. 2010;**5**(3-4):139-188

Chapter 5

Ontologies as a Tool for Formalizing Data Validation Rules

Nicholas Nicholson and Iztok Štotl

Abstract

Comparison of health data across national or even regional boundaries is a challenging task. Data sources, data collection methods, and data quality can vary widely and the quality of the indicators themselves is dependent upon the veracity of the underlying data. For any trans-regional or trans-national comparison of indicators, it is imperative to ensure data are appropriately validated. Ontologies provide a number of functionalities to help in this process. Data rules can be formalized using the ontology axioms, which are useful for removing the ambiguities of rules expressed in natural language. In addition, the axioms serve to identify the metadata and their corresponding semantic relationships, which can in turn be linked to standard data dictionaries or other ontologies. Moreover, ontologies provide the means for encapsulating the underlying data model of the domain allowing the rules and the data model to be maintained in a single application. Finally the expression of the axioms in description logic, as supported for example by the web ontology language, allows machine reasoning to validate data sets automatically against the formalized rules.

Keywords: web ontology language, data harmonization, data validation, data rules, description logic, linked metadata

1. Introduction

Data validation is a key part of the overall data harmonization process that allows meaningful comparison or integration of different data sets. This is particularly important for the derivation of indicators, which may be used for comparison or benchmarking purposes across countries or regions. Prime examples are population-based disease surveillance programs and environmental monitoring and control programs.

Disease monitoring and surveillance is a particular focus of the European Union and a number of pan-European registry networks exist for this purpose. The European Network of Cancer Registries (ENCR) is the most established surveillance network incorporating over 150 separate regional or national registries [1]. A similar initiative in the United States is the Surveillance, Epidemiology, and End Results (SEER) program [2].

In order to help harmonize the data, which may be collected via different processes from different sources, registry networks generally agree a core or common data set that comprises the most accessible, important and well-defined variables. As an example the ENCR common data set consists of about 50 variables [3]. Even though the common data set variables are generally well defined, they may not necessarily be described in a manner that easily allows semantic linkage or cross-reference. Furthermore, they may depend on domain-specific knowledge not readily available to data users outside the domain.

Indictors for comparison purposes tend to be derived from common data sets since they constitute the variables that are the most harmonized within a disease domain. It is particularly important that the underlying data of the indicators are consistent and complete to avoid erroneous conclusions or bias in the results [4]. Ensuring an adequate level of consistency however is quite difficult to achieve in practice given the heterogeneity of data sources and data-collection processes.

Assuming a pre-defined level of quality, data consistency can nevertheless be verified using rule-based systems to check that the individual data fields are present and within the expected ranges. More complex, inter-variable rules check data consistencies between variables and their values. Other consistency checks can compare the frequency of occurrences of specific values of data. All these checks provide greater confidence in the fidelity of data sets for comparison purposes [5].

2. Specification of the rule base

Specifying the data-validation rules in an optimal way is itself a challenge. Rules are often described using natural language which, whilst having the advantage of making them more readable, leads to ambiguities for anything other than the most simple rules. Complex rules with dependencies on multiple variables can be illustrated more easily via a series of tables that constrain the values of the variables not forming the major focus within a particular table. Ensuring the consistency and verifying the accuracy of the rules across multiple tables is not straightforward and leads to considerable maintenance overheads.

The ENCR common data set comprises variables describing a tumor, such as: morphology (type of tumor); behavior (how the tumor acts in the body); topography (organ affected); basis of diagnosis (how the tumor was diagnosed); grade (how the tumor cells compare with normal cells under the microscope); and stage (extent of the tumor). Morphology, behavior, topography, and grade are specified by codes adhering to the international classification of diseases for oncology, edition 3 (ICD-O-3) [6]. Stage for solid tumors is generally specified according to the globally recognized TNM staging system describing the extent of cancer disease, where the "T" component is related to the size of the tumor or its invasion into local structures; the "N" component is related to the number and nature of lymph node groups adjacent to the tumor with evidence of tumor spread; and the "M" component is related to the presence of local or distant metastatic sites. The rule interdependencies of all these tumor-description variables in the ENCR rules are illustrated in **Table 1**. To manage more easily the complexity of the interdependencies, the rules are divided into nine separate sets of tables, namely:

1. age/morphology/topography;

2. sex/topography;

3. sex/morphology;

	Morph	Topog	Age	Sex	BoD	Grade	Beh	Stage
Morph	■	X	X	X	X	X	X	X
Topog	X	■	X	X	X	▨	▨	X
Age	X	X	■	▨	X	▨	▨	X
Sex	X	X	▨	■	X	▨	▨	▨
BoD	X	X	X	X	■	▨	▨	X
Grade	X	▨	▨	▨	▨	■	X	X
Beh	X	▨	▨	▨	▨	X	■	X
Stage	X	X	X	▨	X	X	X	■

Table 1.
Rule interdependencies (marked with an "X") of some of the main variables within the ENCR common data set. Morph = morphology; Topog = topography; BoD = basis of diagnosis; Beh = behavior. The shaded cells indicate no interdependencies.

4. basis of diagnosis/morphology/topography/age;

5. grade/morphology/behavior;

6. morphology/topography;

7. topography/stage-grouping/TNM;

8. topography/topography-grouping (for multiple primary tumor conditions);

9. morphology/morphology-grouping (for multiple primary tumor conditions).

Given the size of the tables, only a few excerpts are shown for illustrative purposes in **Tables 2**–**6**. Whereas they are specific to the ENCR common data set, they are nevertheless indicative of the sorts of difficulties faced by other rule sets defined in a similar fashion.

Apart from the difficulty of ensuring consistency across the rule tables, a further drawback to defining rules in this way relates to the intricacy they impose on compiling a test data set. A comprehensive test data set is important for verifying the ability

Age group (years)	Morphology	Topography
0–2	Hodgkin lymphoma 9650–9667	—
>7	Malignant extra-cranial and extra gonadal germ cell: 9060–9065, 9070–9072, 9080 9085, 9100–9105	C00-C55, C57-C61, C63-C69, C73-C750, C754-C768, C80
0–14	Mesothelial neoplasms: 9050–9053	Any
< 40	Adenocarcinoma: 8140	C61

Table 2.
Unlikely and rare combinations of age and tumor type (excerpt from table 3 in [3]).

Basis of Diagnosis	Morphology (and topography, age, and sex where indicated)
2	8000, 8720, 8800, 8960 (age 0–8), 9140, 9380 (C717), 9384/1, 9500 (age 0–9), 9510 (age 0–5), 9530–9539 (C70), 9590, 9800
4	8000, 8150–8154, 8170, 8270–8281(C751), 9100 (female age 15–49), 9500 (age 0–9), 9732 (and age 40+), 9761 (and age 50+)
6	≠ 8000; 9590–9731; ≠ 9732; ≠ 9733–9760; ≠ 9761; ≠ 9762–9992

Table 3.
Valid combinations for basis of diagnosis and morphology (excerpt from figure 2 in [3]).

Sex	Topography
Female	C60, C61, C62, C63
Male	C51, C52, C53, C54, C55, C56. C57, C58

Table 4.
Invalid combinations for sex and topography (excerpt from table 4 in [3]).

Morphology	Allowed topography	Disallowed topography
8010–8589		C38, C40-C42, C47, C480, C49, C70-C72, C77
8090–8095, 8097, 8100–8103, 8110	C300, C44, C51, C60, C632	
8800–8811, 8814–8831, 8840–8921, 8963, 8990, 8991, 9040–9043, 9120–9150, 9170, 9540, 9550, 9561, 9580, 9581		C420, C421, C77

Table 5.
Morphology codes and allowed/refused topography codes (excerpt from table 8 in [3]).

of data-checking software to trap the different types of errors against the rules. In constructing a test data set, it is necessary to keep record of the variables set incorrectly for each individual test case.

Creating a test record using the tabular rules requires one first to establish a valid morphology/topography combination (one table look-up), then a correct morphology/behavior combination (second table look-up), and thereafter multiple table look-ups for all the other variable interdependencies. Given that not all possible morphology/topography combinations lead to defined combinations of the other variables, it becomes an arduous task to follow this process to completion. In practice, what is done is to start from a real cancer registry data set and systematically set the variables to incorrect values. However, such an approach does not guarantee all possible record combination conditions are thereby tested, potentially leading to undetected bugs in the validation software.

For many practical reasons therefore, a more formal representation of the data rules is necessary. Ontologies are interesting since they provide the basis for doing this in a way that is also integrated with the underlying data model.

Stage	T	N	M
Thyroid gland – papillary or follicular, < 45 years			
I	Any T	Any N	M0
II	Any T	Any N	M1
Thyroid gland – papillary or follicular, ≥ 45 years			
I	T1a, T1b	N0	M0
II	T2	N0	M0
III	T3	N0	M0
	T1, T2, T3	N1a	M0
IVA	T1, T2, T3	N1b	M0
	T4a	N0, N1	M0
IVB	T4b	Any N	M0
IVC	Any T	Any N	M1

Table 6.
TNM edition 7 stage grouping and T, N, M values for thyroid gland (C73) papillary or follicular (excerpt from appendix III in [3]).

3. The relationship between ontologies and description logics

Computational ontologies describe and categorize classes of objects and specify the relationships associated with those classes and categories. This information is captured using axiomatic constructs that provide an appropriate mechanism for describing the majority of the ENCR data rules.

There is in fact a very close relationship between the axiom constructs and description logics (DLs) [7], which are themselves closely related to first-order and modal logics. Since first-order logic draws from a well-established mathematical foundation, DLs provide a solid formal framework for representing axioms that can be developed using the more readily understandable ontology constructs.

DLs form a family of knowledge representation languages that are distinguished by their level of expressivity [8]. Expressivity refers to the expressive power of the language governed by the types of operations it can support. The base language is attributive language (AL) supporting concept intersection (⊓), some level of negation (¬), universal restrictions (∀), and existential restrictions (∃) with limited quantification. The restriction operators ∀ and ∃ are used for qualifying the entities on which a given role acts, with ∃ specifying the notion of an "at-least-one relationship" and ∀ the notion of an "only relationship"; they are similar to the existential and universal quantifiers of first-order logic.

The addition of complex concept negation (C), which includes concept disjunction (⊔), increases the expressivity to attributive language with complements (ALC) that already provides quite a powerful expressivity able to handle many types of data rules. A language of higher expressivity is SHOIN, where S refers to ALC with transitive roles, H to role hierarchy, O to nominals, I to inverse properties, and N to cardinal restrictions. Higher expressivities are also possible but there is a trade-off between expressivity and computational cost for automatic reasoning.

In DL terminology, a knowledge base has two distinct components – a terminological part or TBox, and an assertional part or ABox. An additional term RBox is sometimes used to denote an extended set of role axioms that are described by the letter R in higher expressivities such as SROIQ [8].

The distinction between the TBox and ABox is sometimes also made in the division between ontologies and knowledge graphs [9]. An ontology is considered as a schema that captures the semantic data model using classes, relationships, and attributes (i.e. the TBox, where concepts stand for classes and roles for relationships). A knowledge graph in contrast contains specific instances following the semantic data model represented by the ontology (i.e. the ABox).

3.1 Web ontology language

The World Wide Web Consortium (W3C) describes the web ontology language (OWL) as "a semantic web language designed to represent rich and complex knowledge about things, groups of things, and relations between things". It refers to OWL documents as ontologies [10]. OWL is structured closely along the lines of DLs and provides support for automatic reasoning. It uses the terminology of classes and properties (instead of concepts and roles) for the TBox and represents the ABox as a set of individuals instanced (or asserted) from the TBox axioms.

A number of free, open-source graphical user interface OWL editors are available (e.g. Protégé [11]) that greatly ease the task of ontology development. It is generally more straightforward to define classes and relationships from an ontological point of view than construct them from scratch using DL. The DL expressions can afterwards be determined from the resulting OWL axioms.

4. OWL: A formal framework for the specification of the data rules

OWL's roots in DL allow a formal context to be established for data rules that can overcome the inherent ambiguities associated with their formulation in natural language. Given the relatively rich set of logic operators available however, care is required in deciding how best to formulate the axioms. Unfortunately, there is no simple set of guidelines to help with this task since it is very much dependent on how the ontology will be used. Moreover, DL expressivity comes at the cost of computational speed [12] and where this is important, it is preferable to restrict the DL expressivity to the extent necessary.

4.1 Representation of the data rules

By way of illustration, the following simple examples are only intended to show how some of the rules depicted in **Tables 2–6** can be encoded in DL. With reference to **Table 5** (morphology/topography), capturing the fact that the topography code *C*300 (nasal cavity) with a morphology code of 8090 (basal cell carcinoma) is a permissible combination, one can create an OWL axiom stating that *C*300 is a subclasss of the object property *hasMorphology* with a filler class *M*_8090 (where the prescript *M*_ has been added for convenience to represent morphology). This statement is represented in DL by:

$$C300 \sqsubseteq \exists\, hasMorphology.M\_8090 \qquad (1)$$

In a similar manner, one can capture the rule in the last row of **Table 2** that an ICD-O-3 topography code *C*61 (prostate gland) together with a morphology of 8140 (adenocarcinoma) is unlikely in men aged less than forty years at diagnosis. This rule, which requires use of an OWL data property, can be framed in such a way to say that for a combination of topography and morphology, the expected age of patients is above thirty-nine years:

$$C61 \sqcap M\_8140 \sqsubseteq \exists \, expectedAge.\{>39\} \tag{2}$$

The introduction of another axiom stating that the conjunction of an expected age of more than thirty-nine years and a patient age at diagnosis of less than forty years is an improbable scenario, Eq. (3), would flag a potential coding error (via subsumption under the class *ImprobableAge*) for any prostate tumor cases with morphology code 8140 for patients younger than forty years of age.

$$\exists \, expectedAge.\{>39\} \sqcap patientAgeAtDiagnosis.\{<40\} \sqsubseteq ImprobableAge \tag{3}$$

Clearly such a rule would have to be replicated for all the relevant upper age restrictions provided in the rule table. To avoid logic conflicts, a modified set of axioms would need to be created for the rules with lower age restrictions, c.f. row 2 in **Table 2**.

By building up axioms in this manner, all the rules relevant to a given class or hierarchy of classes can be defined. The advantage is that each rule governing a class of objects is visible on the ontology editor's view of the class, unlike the representation of the rules in **Tables 2–6** where one has to search between various tables to ascertain all the rules pertinent to a particular entity. As observed earlier, this greatly simplifies the task of building up test cases of data both to validate the behavior of the rules as well as to construct comprehensive test data sets.

4.2 Automatic reasoning

Owing to its DL foundations, OWL provides the possibility for automatic reasoning. Automatic reasoning is a valuable tool for detecting rule violations in a set of data records. Eq. (3) provided an example where a reasoner could flag a potential coding error in a cancer case.

In designing error-trapping axioms, it is important to be aware of the issues relating to the open world assumption of DL. The open world assumption holds the view that anything not explicitly stated can only be assumed to be unknown. This is in contrast to the closed world assumption in which anything not explicitly stated is considered incorrect (typical for rules expressed for instance in Datalog). The open world assumption has implications in the subsumption of classes in a hierarchy and can dictate the structure of the ontology dependent on the reasoning requirements.

Data rules, which by definition are prescriptive in the dependencies between data variables, are more suited to the closed world assumption. Axioms may therefore have to be written in such a way that serves to force class subsumption in an otherwise open world view. One means for achieving this is to "invert" the class tree – which may be more easily clarified by the following simple practical example. Say we wished to subsume a class with certain attributes (e.g. a class having a topography code of *C*40 and a morphology with code 919) under a general classification class of *Osteosarcoma*. Following the traditional approach of constructing classes using an ontology editor such as Protégé, we might declare an axiom such as:

$$Osteosarcoma \sqsubseteq C40 \sqcap M\_919 \tag{4}$$

If we were to declare a class *TumorCase* also subclassed from an intersection of $C40$ and M_919 and then run the reasoner, we would find that our *TumorCase* class had not been classified under (i.e. subsumed by) the class *Osteosarcoma*. This is due to the open world assumption since it cannot be assumed that the class *Osteosarcoma* is not subclassed from other classes that have not been explicitly stated. It cannot therefore be assumed that the *TumorCase* class is contained by the *Osteosarcoma* class – there is not enough information to say.

The problem can be circumvented either by creating an equivalence (using defined classes) or by inverting the subclass definition. Creating many equivalences with complex classes can however lead to unintended consequences. For example, if the containment operator ($\sqsubseteq$) in Eq. (3) were to be replaced by an equivalence ($\equiv$), and if this approach were to be replicated for the whole set of axioms modeling each of the age-restricted rules (c.f. **Table 3**), then all the expressions on the left-hand-side of the equivalence would also become equivalent (since they are all equivalent to the class *ImprobableAge*) and this would be erroneous. Alternatively, the subclass definition of Eq. (4) can be inverted as indicated in Eq. (5):

$$C40 \sqcap M\_919 \sqsubseteq Osteosarcoma \tag{5}$$

Running the reasoner now would result in the subsumption of the class *TumorCase*. under the class *Osteosarcoma*.

This method of axiom formulation has been coined “being complex on the left-hand side” [13]. Ontology editors such a Protégé lead developers to put the complexity on the right-hand side of the class containment relation (i.e. subclassing from complex classes). Although moving the complexity to the left-hand side can overcome the subsumption issues of the open world view, it tends to obfuscate the ontology structure. Eq. (3) is a further example of defining axioms following this approach.

Regarding the different formulations for expressing the rule illustrated in Eqs. (4) and (5), it is instructive to note that the equivalence expression:

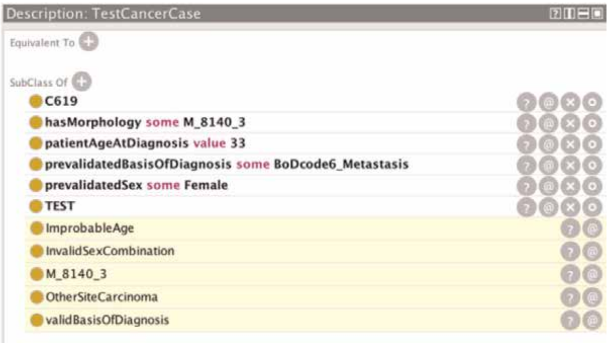
$$C40 \sqcap M\_919 \equiv Osteosarcoma \tag{6}$$

is in fact a short-hand way of writing the implied DL expression:

$$C40 \sqcap M\_919 \sqsubseteq Osteosarcoma, Osteosarcoma \sqsubseteq C40 \sqcap M\_919 \tag{7}$$

Figure 1 is a view from the Protégé application showing the result of reasoning based on the classes and properties given in an imaginary cancer test case. The non-highlighted lines indicate the information passed into the reasoner and the lines highlighted with yellow background show the extra information returned by the reasoner. Noting that the topography class $C619$ is a subclass of $C61$ and the morphology class M_8140_3 is a subclass of M_8140, and in accordance with the rules provided in **Table 2** (row 4) and **Table 3** (row 3), and **Table 4** (row 1), the reasoner has ascertained that: the age at diagnosis is improbable for the morphology/topography combination; the basis of diagnosis is correct; and the combination of sex and topography is incorrect. The question mark in the gray circle on the highlighted lines provides the means of polling the reasoner to understand why it has subsumed the class under the identified class.

Protégé also provides a graphical view on the inferred classification tree for the named classes (unnamed classes are not visible). **Figure 2** provides an amplification of the classification tree summarized in **Figure 1**.



Description: TestCancerCase

Equivalent To

SubClass Of

- C619
- hasMorphology some M_8140_3
- patientAgeAtDiagnosis value 33
- prevalidatedBasisOfDiagnosis some BoDcode6_Metastasis
- prevalidatedSex some Female
- TEST
- ImprobableAge
- InvalidSexCombination
- M_8140_3
- OtherSiteCarcinoma
- validBasisOfDiagnosis

Figure 1.
Information added from the reasoning process (highlighted lines) based on the prior information of classes asserted in a test case (non-highlighted lines).

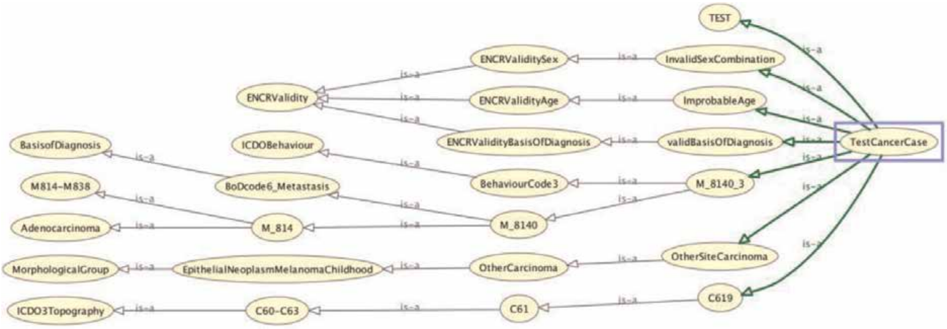
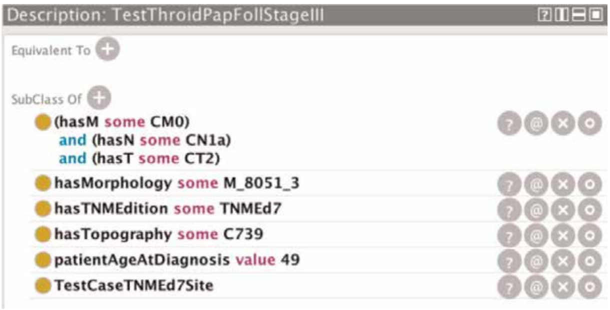


Figure 2.
Graphical view of the classification structure (containing both asserted and inferred classes) of the cancer test case shown in ***Figure 1****.*



Description: TestThroidPapFollStageIII

Equivalent To

SubClass Of

- (hasM some CM0)
  and (hasN some CN1a)
  and (hasT some CT2)
- hasMorphology some M_8051_3
- hasTNMEdition some TNMEd7
- hasTopography some C739
- patientAgeAtDiagnosis value 49
- TestCaseTNMEd7Site

Figure 3.
Thyroid cancer TNM test case to verify the class subsumption results from the reasoner.

The reasoner can be polled to understand the reasoning applied for class subsumption. **Figure 3** shows a cancer test case for the thyroid gland (*C*739) restricted to TNM information to check whether the test case is subsumed under stage III (c.f. **Table 6**, row 7b). **Figure 4** is the classification tree resulting from the automatic reasoning process on the TNM test case of **Figure 3**. It can be seen that the reasoner has correctly subsumed the test case under the stage III class." **Figure 5** shows the results from polling the reasoner to understand why it subsumed the test class under

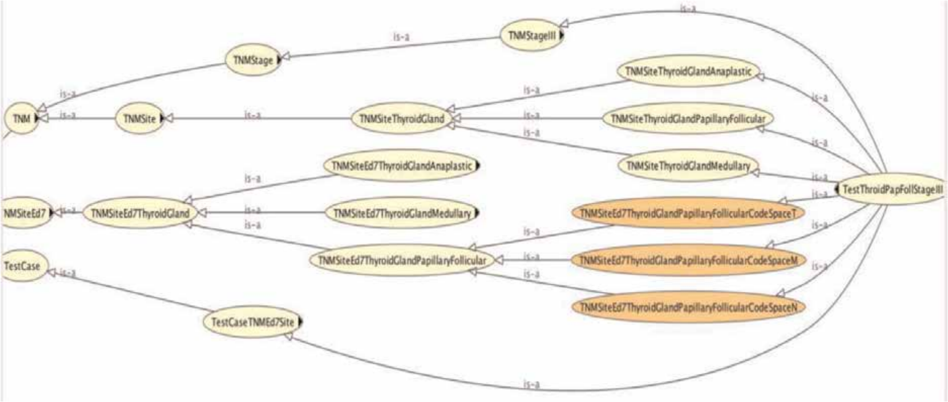
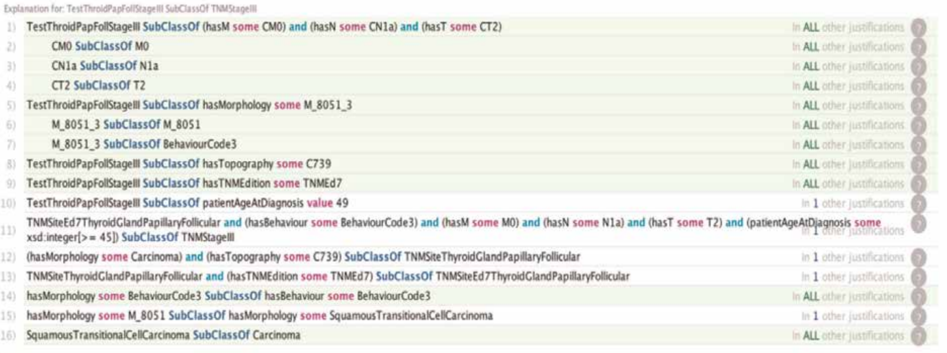

TNMStageIII
TNMStage
TNMSiteThyroidGlandAnaplastic
TNM
TNMSite
TNMSiteThyroidGland
TNMSiteThyroidGlandPapillaryFollicular
TNMSiteThyroidGlandMedullary
TNMSiteEd7ThyroidGlandAnaplastic
NMSiteEd7
TNMSiteEd7ThyroidGland
TNMSiteEd7ThyroidGlandMedullary
TestThroidPapFollStageIII
TNMSiteEd7ThyroidGlandPapillaryFollicularCodeSpaceT
TNMSiteEd7ThyroidGlandPapillaryFollicular
TNMSiteEd7ThyroidGlandPapillaryFollicularCodeSpaceM
TestCase
TNMSiteEd7ThyroidGlandPapillaryFollicularCodeSpaceN
TestCaseTNMEd7Site
is-a

Figure 4.
*Classification tree of the thyroid cancer test case of **Figure 3**, showing that the reasoner has correctly identified the stage III class (the top-most class in the figure) as required from the rule table shown in **Table 6**. The classes shaded in the darker color represent defined classes (classes with some equivalence conditions).*



```
Explanation for: TestThroidPapFollStageIII SubClassOf TNMStageIII
1)  TestThroidPapFollStageIII SubClassOf (hasM some CM0) and (hasN some CN1a) and (hasT some CT2)                      In ALL other justifications
2)      CM0 SubClassOf M0                                                                                              In ALL other justifications
3)      CN1a SubClassOf N1a                                                                                            In ALL other justifications
4)      CT2 SubClassOf T2                                                                                              In ALL other justifications
5)  TestThroidPapFollStageIII SubClassOf hasMorphology some M_8051_3                                                   In ALL other justifications
6)      M_8051_3 SubClassOf M_8051                                                                                     In ALL other justifications
7)      M_8051_3 SubClassOf BehaviourCode3                                                                             In ALL other justifications
8)  TestThroidPapFollStageIII SubClassOf hasTopography some C739                                                       In ALL other justifications
9)  TestThroidPapFollStageIII SubClassOf hasTNMEdition some TNMEd7                                                     In ALL other justifications
10) TestThroidPapFollStageIII SubClassOf patientAgeAtDiagnosis value 49                                                In 1 other justifications
11) TNMSiteEd7ThyroidGlandPapillaryFollicular and (hasBehaviour some BehaviourCode3) and (hasM some M0) and (hasN some N1a) and (hasT some T2) and (patientAgeAtDiagnosis some
    xsd:integer[>= 45]) SubClassOf TNMStageIII                                                                         In 1 other justifications
12) (hasMorphology some Carcinoma) and (hasTopography some C739) SubClassOf TNMSiteThyroidGlandPapillaryFollicular      In 1 other justifications
13) TNMSiteThyroidGlandPapillaryFollicular and (hasTNMEdition some TNMEd7) SubClassOf TNMSiteEd7ThyroidGlandPapillaryFollicular   In 1 other justifications
14) hasMorphology some BehaviourCode3 SubClassOf hasBehaviour some BehaviourCode3                                      In ALL other justifications
15) hasMorphology some M_8051 SubClassOf hasMorphology some SquamousTransitionalCellCarcinoma                          In 1 other justifications
16) SquamousTransitionalCellCarcinoma SubClassOf Carcinoma                                                             In ALL other justifications
```

Figure 5.
Reasoner justification for the subsumption of the thyroid cancer test case under the TNMStageIII class.

the *TNMStageIII* class. The specific rule is stated in line 11 of the figure and the other lines provide the reasons for subsuming the classes asserted in the test case under the various classes in the rule itself.

Automatic reasoning can be performed using both TBox axioms and ABox axioms. Since data rules are more often associated with classes of objects, TBox reasoning is in most cases sufficient and can reduce computational costs. Most of the ENCR data rules can be modeled by TBox axioms apart from those, for example, that pertain to multiple tumors (where a person has more than one type of cancer). The rules specify the topography and morphology combinations of any two tumors to be considered different and since two entities with the same class attributes have to be compared, the use of ABox axioms is necessary. Modeling of the multiple primary rules on the basis of DLs supported by OWL has been addressed at length in [14].

The ability to include closed world reasoning in OWL would be ideal and has been made possible to a certain degree via the incorporation of the semantic web rule language (SWRL) into the semantic web stack. SWRL is based on first-order Horn-logic in which rules in Datalog are also expressed [15], but requires an ABox. Another expressive logic formalism allowing some integration of open- and closed-world reasoning is minimal knowledge and negation as failure (MKNF) [16]. This formalism is being developed in a unifying framework in the KAON2 infrastructure [17].

4.3 Encapsulation of the data model

The axiomatic constructs of an ontology are useful for capturing many of the different aspects of a data model that for relational database models have traditionally been divided across three independent levels of abstraction. Namely, the conceptual schema (describing the semantics of the domain and the scope of the model); the logical schema (describing the structure of the information, as for example a relational database schema); and the physical schema (describing the physical means of storing the data) [18].

One of the strengths of OWL is its relationship with the resource description framework (RDF), which serves as the data interchange layer of the semantic web stack [19]. RDF data is in essence a network of connected triplets of resources, in which the resources at the edges of the triplets (subject and object) are related by the resource in the middle of the triplet (predicate). Each resource is identified by a uniform resource identifier (URI). All OWL constructs are described in terms of RDF data, allowing ontologies to bridge the traditional divide between conceptual and logical levels of abstraction and providing a richer, more integrated data model description framework.

The flexibility and descriptive power of an ontology present their own sets of challenges however. While the usefulness of ontologies is widely acknowledged, the task of building a good ontology is a particularly hard one and falls within the developing domain of ontology engineering [20]. Designing an appropriate ontology does not only depend on a thorough understanding of the domain to be modeled, but must be performed circumspectly in view of the ontology's purpose and future extensibility. There are pitfalls in making an ontology too granular or not granular enough – the result is either a multiplication of application-specific ontologies that cannot easily be integrated, or an ontology overly generic to be useful to any particular application. OWL provides the functionality for importing ontologies that allows larger ontologies to be built up in a modular fashion and this can aid the design process if performed carefully [21].

There are also certain design aspects to take into account that can affect the overall structure of the ontology. One important consideration relates to the extent to which the ontology is to be used in a pre-coordinated or post-coordinated way [22]. Pre-coordination refers to the situation in which all the terms and relationships are stated explicitly in the axioms and leads to a static use of the ontology, whereas post-coordination refers to the more dynamic situation in which new relationships are determined by the automatic reasoning process on the basis of the pre defined axioms. The pitfalls are exacerbated in applications that need to tweak the normal approach to structuring class hierarchies to overcome restrictions in post-coordination that the open world assumption places on class subsumption.

If the axioms describing the data rules are developed circumspectly however, the advantage is that the data model falls out almost by default – the data rules necessarily identify all the concepts within the domain as well as their inter-relations. This may require an iterative process combining both the bottom-up approach of developing axioms in DL and the top-down approach of structuring the ontology, while testing each stage of the development with the reasoner.

The task of developing a data model in an ontology used in a predominantly pre-coordinated way is perhaps more straightforward and does not require too much juggling in defining the axioms. Moreover, the axioms can be constructed in the more usual manner of subclassing from complex classes. The intelligence of validating data sets would however need to be moved from the ontology to a computer program (for instance via the OWL-API) thereby compounding maintenance issues. The advantage of encapsulating the intelligence in the ontology is that all the knowledge is contained in one application and maintenance aspects are thereby confined to that one application.

4.4 Metadata by default

Elements in an ontology are described in terms of their semantic relations to other elements in the ontology thereby providing a description and context, or in other words the metadata, of the element. Moreover, since each element in an OWL ontology is uniquely defined by a uniform resource identifier (URI), it is readily linkable with other web resources. This allows any element to be associated with other relevant resources via linked open data (LOD) principles. Using knowledge organization schemes, such as simple knowledge organization system (SKOS), it becomes a straightforward matter to link OWL resources semantically with other web-based resources such as data-dictionary or thesauri elements.

The interlinking of any OWL resource to other web resources, especially to other RDF resources, provides a powerful and extensible means of capturing all the necessary metadata components for comprehensively describing a data model element. This aspect has been exploited to create extensive frameworks of distributed metadata registries that allow the reuse of existing metadata resources [23].

It is important to emphasize that a number of complementary tools exist that can be used together to provide a more comprehensive toolkit for validating different types of data rules. Included in the semantic web standards are the shape languages: shape expression (ShEx) and shapes constraint language (SHACL) for providing structural schema for RDF data. There are also additional tools for polling knowledge bases such as the SPARQL protocol and RDF query language (SPARQL) as well as those for extending the expressivity of OWL DLs, such as SWRL. Depending on the type of rule, some of these tools may be more suitable than others; however, since they are agreed or proposed semantic web standards and based on the standard model for data interchange (RDF), they can all reference the elements of a data model described in RDF. This provides a highly flexible and versatile environment in which to develop an integrated toolkit. **Table 7** gives a summary breakdown of these applications with the sorts of operations they support and the components of a knowledge base to which they are applicable.

Whereas other tools and languages (e.g. Datalog) are also available for validating data, and may arguably be more appropriate for defining rules predominantly based on closed-work scenarios, they fall down in this aspect of unifying the rules with the data model and the metadata, especially in the LOD sense. For federated data-validation processes, the unification of all these elements brings many advantages in terms of data linkage, maintenance, and collaborative development. Having said that, OWL is not able to handle all types of validation checks – such as those for example requiring comparison of dates, checking of frequencies of occurrence, or expressing certain relations between individuals. ShEx, SWRL, and SPARQL can all go some way to handling such checks. SWRL and SPARQL however require an ABox and SWRL has implications on decidability [24]. Moreover, introducing an ABox can create

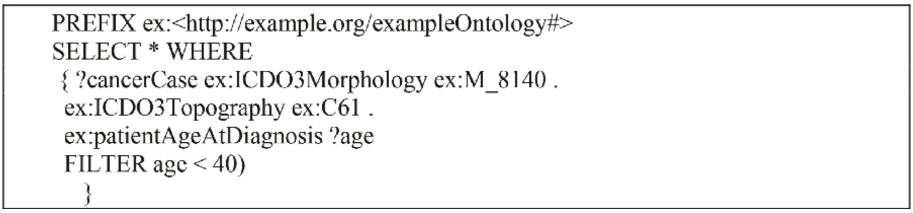
Application	Scope	Inference mechanism	Types of operations supported	TBox, ABox focus
OWL	Knowledge bases, DL	Yes	Complex inter-variable checks supported by DL expressivity	TBox, Abox
SWRL	Extension of logic to OWL	Yes	Complex inter-variable checks	ABox
OWL-API	Programming interface to OWL ontologies	Yes	Complex inter-variable checks supported by DL expressivity and additional computer logic	TBox, Abox
ShEx	Grammar check of RDF graphs	No, although can be used in post-coordination	Ensuring RDF data conforms to an expected template. Can perform some inter-variable dependency checks and verify if values of variables are in range.	TBox, Abox
SHACL	Constraint requirements of RDF graphs	Some	Ensuring RDF data conforms to a given set of constraints: Can compare date fields. More suitable for validating RDF graphs than conformance with a specific template (for which ShEx is better)	TBox, Abox
SPARQL	Query language for RDF data	No, although can be used in post-coordination	Querying of data by user-defined query-language constructs	Abox

Table 7.
Summary breakdown of some of the semantic web standard applications with the sorts of operations they support.

performance issues for DL reasoning when many hundreds of thousands of individuals are involved and requires careful consideration in the ontology design phase. An alternative is to create an ABox and use SPARQL querying instead of DL reasoning but this would move the rule logic out of the ontology and into the SPARQL query scripts.

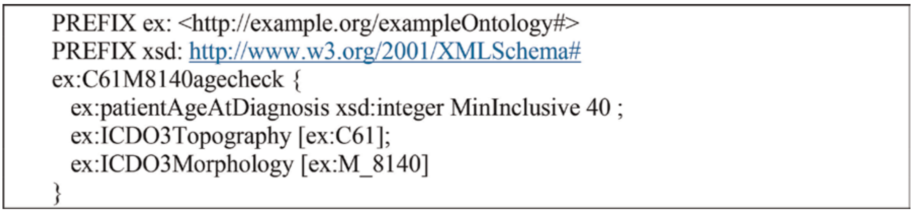
An example for handling the axioms of Eqs. (2) and (3) using a simple SPARQLscript to list all the associated erroneous cancer-case records is shown in **Figure 6**. A ShEx script for checking the same condition is shown in **Figure 7**. The same rule using SWRL could be expressed as shown in **Figure 8**.

The effort required to maintain the rule base developed with such tools however would be considerable and it would make more sense to use them in a pre-processing stage on the data to be validated (translated beforehand into RDF) for those types of checks that cannot be handled within the ontology itself. ShEx in particular provides a valuable pre-processing tool to check the ranges and formats of variables.



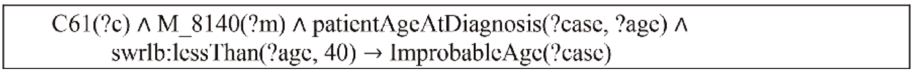
```
PREFIX ex:<http://example.org/exampleOntology#>
SELECT * WHERE
 { ?cancerCase ex:ICDO3Morphology ex:M_8140 .
  ex:ICDO3Topography ex:C61 .
  ex:patientAgeAtDiagnosis ?age
  FILTER agc < 40)
     }
```

Figure 6.
An example of a SPARQL script to list all the erroneous patient-age related cancer-case records associated with a particular combination of topography and morphology codes.



```
PREFIX ex: <http://example.org/exampleOntology#>
PREFIX xsd: http://www.w3.org/2001/XMLSchema#
ex:C61M8140agecheck {
  ex:patientAgeAtDiagnosis xsd:integer MinInclusive 40 ;
  ex:ICDO3Topography [ex:C61];
  ex:ICDO3Morphology [ex:M_8140]
}
```

Figure 7.
An example of a ShEx script to trap any erroneous patient-age related cancer-case records associated with a particular combination of topography and morphology codes.



```
C61(?c) ∧ M_8140(?m) ∧ patientAgeAtDiagnosis(?case, ?age) ∧
        swrlb:lessThan(?age, 40) → ImprobableAge(?case)
```

Figure 8.
An example of an SWRL rule to catch the same validation errors as for ***Figures 6*** *and* ***7****.*

5. Role of ontologies in data harmonization

The focus until this point has been on how ontologies can provide many advantages in the task of data validation against a set of specific data-validation rules. Checking the conformity of data against such rules is just one element in the whole process of data harmonization.

Data harmonization is a term that eludes a clear and concise definition, perhaps partly due to its dependence on the context to which it is applied [25, 26] as well as the fact that it is a multistep activity involving both technical and social processes [5, 26]. An idealized breakdown of these steps has been provided in [5] based on the accumulated experience gained by the Comprehensive Center for the Advancement of Scientific Strategies (COMPASS) resulting from multiple data-harmonization projects across widely different types of data, collaborators, and scientific questions. Whereas not all projects were found to follow all steps and the order of the steps might vary, the six most common steps identified were:

1. Identification of the questions that the harmonized data set is required to answer

2. Identification of the high-level data concepts required to answer those questions

3. Assessment of the data availability for the data concepts

4. Development of CDEs for each data concept

5. Mapping and transformation of individual data points to CDEs

6. Quality-control procedures

In this breakdown, the process of data validation falls manly under steps 5) and 6) although it should be stressed that validation forms only part of the quality-control procedures of step 6). Other fundamental quality metrics consist of the following dimensions: completeness, consistency, accuracy, timeliness, uniqueness, and auditability [27]. Moreover, different entities in the data process may be responsible for ensuring the quality of the data associated with these separate dimensions. They are nevertheless all important for ensuing an appropriate level of harmonization that allows meaningful comparison or integration of data and it would not be correct to state that data solely validated against a set of validation rules have the prerequisite level of quality for purposes of data comparison.

The degree to which data are harmonized depends ultimately on the specific end use, but the step can never entirely be ignored. In the field of health for example, data harmonization is a critical step in pooling data sets for increasing the power of individual epidemiological studies [5]. It is also a necessary part of health management decision-making, particularly with regard to: clinical decision-making for individual patient clinical management or clinical support and quality improvement tools; operational and strategic decision-making for health system managers and policy-makers; and population-level decision-making for disease surveillance and outbreak management [26].

The point is that ontologies can play an important part in all stages of data harmonization. Starting from the highest levels of abstraction in the six-step harmonization process presented above, ontologies provide the means to capture and organize the high-level data concepts needed to address the questions the harmonized data are required to answer. Ontologies would moreover be able to formalize the questions in direct reference to the high-level data concepts and help identify any missing concepts as well as to verify the underlying logic of those relationships. The next steps are to identify the availability of the data and to develop common data elements (CDEs). The data may be in an unstructured format. The development of CDEs is a process of structuring the data and the semantic relations described in a domain ontology can help identify the relevant information. The role of ontologies in ETL (extract, transform, load) processes has been extensively reviewed in [28]. In particular, the authors point to the efficacy of ontologies: (a) to formalize the needs and requirements of users and resolve semantic ambiguity; (b) to discover concepts and their relationships; (c) to enrich source data, provide mappings (also generating them automatically) and increase ETL performance and efficiency; and (d) to support configuration and instantiation of ETL patterns. Moreover, the validation rule base for the data can itself be derived automatically from the data themselves using ontological methods [29] and allows a verification of any pre-defined set of validation rules.

6. Conclusions

Data validation is an essential step in the task of ascertaining the veracity and homogeneity of data for data comparison purposes. In the case of structured data, validation is often performed using a set of data validation rules. Using the ontology layer (OWL) of the semantic web stack to perform this task brings a number of major advantages. First, it provides the means of formalizing the rules in DL, thereby removing the ambiguities and redundancies inherent in natural language. Second, it helps encapsulate the data model and integrate the conceptual and logical schemas that have traditionally been separated. The encapsulation of the data model and the definition of the rules in DL is a mutually supportive step that allows the integration of a bottom-up approach (rule definitions) with a top-down approach (classification and semantic context), from which the data model is the result. Third, the data model expressed in OWL automatically incorporates the metadata. All named entities (classes, properties, and individuals) have their own URIs that can be accessed and linked individually. Accessing an OWL link provides the whole semantic context of the entity, which may in turn be annotated with links to other semantic resources to enrich further the contextual information. Other advantages include the possibility of reasoning on the ontology, allowing inferences to be made automatically and providing other semantic relations not explicitly stated a priori in the ontology. Ontologies can also play an important role in more general data harmonization steps. In particular, they can help in defining and formalizing user needs, discovering semantic contexts in unstructured data, and generating semantic mappings.

Whereas ontologies do suffer some drawbacks (such as issues relating to the open world assumption), the fact they can to a large extent unify the underlying data model with the data rules, as well as capture the metadata that can be linked semantically to other metadata dictionaries and classification schemes, makes them an interesting solution. These considerations are of particular importance for applications that need to harmonize data across multiple data providers and heterogeneous data-collection procedures, as well as for improved contextualization of the data that is useful for downstream processes.

Acknowledgements

This work was partly conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

Conflict of interest

The authors declare no conflict of interest.

Nomenclature

CDE	common data element
COMPASS	Comprehensive Center for the Advancement of Scientific Strategies
DL	description logic
AL, ALC, SHOIN, SROIQ	DL expressivities, where:
AL	attributive language
ALC	AL with complements
S	ALC with transitive roles
H	role hierarchy
O	nominals
I	inverse properties
N	cardinal restrictions
R	extended set of role axioms
Q	qualified cardinality restrictions
ENCR	European Network of Cancer Registries
ETL	Extract, Transform, Load
ICD-O-3	International Classification of Diseases for Oncology, third edition

KB	knowledge base
ABox	assertional part of a KB
TBox	terminological part of a KB
RBox	extended set of role axioms in a KB
LOD	linked open data
MKNF	minimal knowledge and negation as failure
RDF	resource description framework
SKOS	simple knowledge organization system
SWRL	semantic web rule language
SHACL	shapes constraint language
ShEx	shape expressions
SPARQL	SPARQL protocol and RDF query language
TNM	TNM classification of malignant tumors
T	size of tumor
N	involvement of regional lymph nodes
M	presence of distant metastasis
URI	uniform resource identifier
OWL	web ontology language
OWL-API	web ontology language application program interface
W3C	World Wide Web Consortium

Author details

Nicholas Nicholson[1]* and Iztok Štotl[2]

1 European Commission Joint Research Centre, Ispra, Italy

2 Department of Endocrinology, Diabetes and Metabolic Diseases, University Medical Centre Ljubljana, Ljubljana, Slovenia

*Address all correspondence to: nicholas.nicholson@ec.europa.eu

References

[1] European Network of Cancer Registries (ENCR). Available from: https://www.encr.eu/ [Accessed: December 26, 2022]

[2] National Cancer Institute. Surveillance, Epidemiology, and End Results Program (SEER). Available from: https://seer.cancer.gov/ [Accessed: December 26, 2022]

[3] Martos C, Crocetti E, Visser O, Rous B, Giusti F. A proposal on cancer data quality checks: one common procedure for European cancer registries. JRC Technical Report, p. 1-99. DOI: 10.2760/429053

[4] Tijhuis M, Finger JD, Slobbe L, Sund R, Tolonen H. In Verschuuren M, van Oers H, editors. Population Health Monitoring. Climbing the Information Pyramid. Cham: Springer; 2019. p. 59-81. DOI: 10.1007/978-3-319-76562-4_4

[5] Rolland B, Reid S, Stelling D, Warnick G, Thornquist M, Feng Z, et al. Toward rigorous data harmonization in cancer epidemiology research: One approach. American Journal of Epidemiology. 2015;**182**(12):1033-1038. DOI: 10.1093/aje/kwv133

[6] World Health Organization. International Classification of Diseases for Oncology (ICD-O) – 3rd Edition, 1st Revision. 2013. Available online: https://apps.who.int/iris/handle/10665/96612 [Accessed: December 26, 2022]

[7] Calvanese D, Guarino N. Ontologies and description logics. Intelligenza Artificiale. 2006;**3**:21-27

[8] Baader F, Horrocks I, Lutz C, Sattler U. An Introduction to Description Logic. Cambridge: Cambridge University Press; 2017. DOI: 10.1017/9781139025355

[9] Schrader B. Enterprise Knowledge. White paper: What's the Difference Between an Ontology and a Knowledge Graph? 2020. Available from: https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/ [Accessed: December 26, 2022]

[10] W3C. Web Ontology Language (OWL). 2012. Available from: https://www.w3.org/OWL/ [Accessed: December 26, 2022]

[11] Protégé. A Free, Open-Source Ontology Editor and Framework for Building Intelligent Systems. Available from: https://protege.stanford.edu/ [Accessed: December 26, 2022]

[12] Calvanese D, De Giacomo G, Lembo D, Lenzerini M, Rosati R. Data complexity of query answering in description logics. Artificial Intelligence. 2013;**195**:335-360. DOI: 10.1016/j.artint.2012.10.003

[13] Sattler U, Stevens R. Being complex on the left-hand side: General concept inclusions. Ontogenesis. 2012. Available from: http://ontogenesis.knowledgeblog.org/1288 [Accessed: December 26, 2022]

[14] Nicholson NC, Giusti F, Bettio M, Negrao Carvalho R, Dimitrova N, Dyba T, et al. An ontology to model the international rules for multiple primary malignant tumours in cancer registration. Applied Sciences. 2021;**11**:7233. DOI: 10.3390/app11167233

[15] Krötzsch M, Rudolph S, Schmitt PH. On the semantic relationship between Datalog and description logics. In: Hitzler P, Lukasiewicz T, editors. Web Reasoning and Rule Systems. RR 2010. Lecture Notes in Computer Science. Vol. 6333. Berlin, Heidelberg: Springer; 2010.

pp. 88-102. DOI: 10.1007/978-3-642-15918-3_8

[16] Motik B, Rosati R. Closing Semantic Web Ontologies. 2006. Available from: http://www.cs.ox.ac.uk/boris.motik/pubs/mr06closing-report.pdf [Accessed: January 10, 2023]

[17] KAON2. Available from: http://kaon2.semanticweb.org/ [Accessed: January 10, 2023]

[18] TopQuadrant. Ontologies and Data Models – are They the Same? 2011. Available from: https://topquadrantblog.blogspot.com/2011/09/ontologies-and-data-models-are-they.html [Accessed: December 26, 2022]

[19] W3C. Resource Description Framework (RDF). 2014. Available from: https://www.w3.org/RDF/ [Accessed: December 26, 2022]

[20] Mizoguchi R. Ontology engineering environments. In: Staab S, Studer R, editors. Handbook on Ontologies. International Handbooks on Information Systems. Berlin, Heidelberg: Springer; 2004. pp. 275-295. DOI: 10.1007/978-3-540-24750-0_14

[21] Cuenca Grau B, Horrocks I, Kazakov Y. Modular reuse of ontologies: Theory and practice. Journal of Artificial Intelligence Research. 2008;**31**:273-318. DOI: 10.1613/jair.2375

[22] Stevens R, Sattler U. Post-coordination: Making things up as you go along. Ontogenesis. 2013. Available from: http://ontogenesis.knowledgeblog.org/1305 [Accessed: December 26, 2022]

[23] Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. Journal of Biomedical Informatics. 2013;**46**:784-794. DOI: 10.1016/j.jbi.2013.05.009

[24] Hitzler P, Krötzsch M, Rudolph S. Knowledge Representation for the Semantic Web Part II: Rules for OWL, KI 2009 Paderborn; Integrationszentrum, Kreis Paderborn; 2009. p. 8-14. Available from: https://www.semantic-web-book.org/w/images/5/5e/KI09-OWL-Rules-2.pdf [Accessed: February 21, 2023]

[25] Paquette J. The Many Marvelous Meanings of "Data Harmonization". Towards Data Science. Canada: Towards Data Science Inc.; 2021. Available from: https://towardsdatascience.com/about-towards-data-science-d691af11cc2f [Accessed: November 16, 2022]

[26] Schmidt BM, Colvin CJ, Hohlfeld A, Leon N. Definitions, components and processes of data harmonisation in healthcare: A scoping review. BMC Medical Informatics and Decision Making. 2020;**20**(1):222. DOI: 10.1186/s12911-020-01218-7

[27] Nicholson N, Giusti F, Neamtiu L, Randi G, Dyba T, Bettio M, et al. Dotting the "i" of interoperability in FAIR cancer-registry data sets. In: Kais G, Hamdi Y, editors. Cancer Bioinformatics [Internet]. London: IntechOpen; 2021. pp. 131-156. Available from: https://www.intechopen.com/chapters/79580. DOI: 10.5772/intechopen.101330

[28] Lorvão Antunes A, Cardoso E, Barateiro J. Incorporation of ontologies in data warehouse/business intelligence systems - a systematic literature review. International Journal of Information Management Data Insights. 2022;**2**(2):100131. DOI: 10.1016/j.jjimei.2022.100131

[29] Brüggemann S, Aden T. Ontology based data validation and cleaning: Restructuring operations for ontology

maintenance. In: Koschke R, Herzog O, Rödiger K-H, Ronthaler M, editors. Informatik 2007 – Informatik trifft Logistik – Band 1. Bonn: Gesellschaft für Informatik e.V.; 2007. p. 207-211. Available from: https://dl.gi.de/handle/20.500.12116/22581 [Accessed: January 10, 2023]

Chapter 6

Reconfigurable Platform Pre-Processing MAC Unit Design: For Image Processing Core Architecture in Restoration Applications

G.N. Chiranjeevi and Subhash Kulkarni

Abstract

The overwhelming majority of image processing algorithms are two-dimensional (2D) and, as a result, their scope is limited. As a result, the 2D convolution function has important implications for image processing needs. The 2D convolution and MAC design processes are used to perform image analysis tasks such as image blurring, softening, feature extraction, and image classification. This study's primary goal is to develop a more efficient MAC control block-based architectural style for two-dimensional convolutions. In image processing applications, convolution deployment, the recommended 2D convolution architectural methodology, is significantly faster and requires far fewer hardware resources. The resulting convolution values are stored in memory when the convolution procedure is completed.

Keywords: pre-processing block, multi-byte fetching, boundary padding, block memory access, Kernel architecture 2D convolution, MAC, image processing FPGA

1. Introduction

Two-dimensional (2D) convolution is used in image and video processing applications to cover a wide range of applications, particularly image smoothing, image sharpening, edge detection, and a few image restoration techniques. All of these platforms require the use of a 2D convolution block in digital image processing. One of the most important procedures in image processing applications is the MAC design procedure. The implementation of a 2D convolution framework involves several steps, and each pixel value has its own meaning. The first step in creating a picture from any input is to convert it into image pixels and store them in the matrix. The user-selected MAC operations and matrices are then executed.

When given specific types of data as inputs, image processing algorithms/architecture [1, 2] perform best the majority of the time. In the vast majority of cases, however, the input image fails to meet critical requirements. Preprocessing occurs prior to application-specific processing [3]. The storage of images is a significant issue in image processing. Over the years, many image file formats have been developed with the goal of representing images in a streamlined and premium manner that can be used on a variety of platforms [4]. Different images of the same type can have a different scale of signal intensities based on preprocessing.

2. Image processing blocks in preprocessing blocks

Preprocessing systems are [5–10] used to prepare data for use in the analysis. The next block of data is now available for optimal use. The main purpose of preprocessing is to improve the quality of the image so that it can be analyzed correctly [11, 12]. Preprocessing can remove unnecessary distortions from a data set and sometimes improve the features that are important for the application you are working with. Some of the features that can be used to improve images depend on the application being used (e.g., general image processing, image enhancement, or image analysis) [13–19], so they can be effective for use by other types of algorithms.

One of the most distinguishing features of the architects who create images based on Vedic computing is the immense amount of information required to create them. To display a grayscale image at a reasonable resolution, 2106 bits are required. Digital images can be stored and transmitted using XSG blocks, but they need to be compressed and edge-detected first. This is done with image compression and edge detection hardware.

In FPGA preprocessing subsystems (DSP modules) [20], protocols are transformed from standard software-compatible representations [21] to more hardware-compatible representations that exploit data parallelization [22, 23] in certain hardware architectures. Signal and Video Processing Framework [24], which often deviates from conventional von Neumann models such as Dataflow [25].

3. Block hardware architecture for preprocessing

This architecture is designed to make image processing easier by prepping data and preparing images which is shown in **Figure 1**. The strategy proposed in this research is organized into five phases: (1) data collection and analysis, (2) model development, (3) testing and validation, (4) implementation, and (5) evaluation. In the data collection and analysis phase, researchers will collect data from a sample of participants to create a database of information. In the model development phase, the researchers will develop a model to explain the relationship between variables. In the testing and validation phase, the model will be tested against data from a different sample of participants to see if it is accurate. The implementation phase will involve the implementation of the model on a large scale, and the evaluation phase will involve assessing the effectiveness of the model. This chapter introduces a preprocessing method for images, as well as an algorithm for automatically screening images using the preprocessing method.

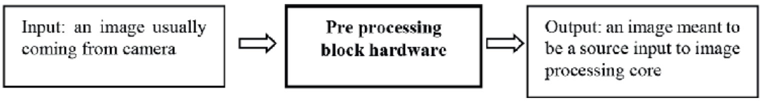

Input: an image usually coming from camera
Pre processing block hardware
Output: an image meant to be a source input to image processing core

Figure 1.
Architecture of the work.

4. Different modes of operations for block memory preprocessing

Two constraints must be considered when designing preprocessing block memory from the user's perspective. The first is to write one pixel value of one byte into the target device's memory at each clock cycle. The representation of the input image (e.g., 512*512) is the second factor to consider. The main goal of this block memory is to have the ability to change the size of the kernel during read and write operations.

When compared to the inbuilt IP core model, the ability to choose the kernel size during read operations is seen as a benefit. During read operations, the IP core can access data in terms of 2 powers bits (i.e., 2, 4, 8, 16, 32, 64), but in the proposed design, it can access data based on kernel requirements and not reserved to any specific values. The observational and experimental during read operation is a general 8 bit pixel value * kernel size.

4.1 Write mode

With during the clock cycle, one pixel of data is written into memory at the address pointed to by the address pointer.

4.2 Read mode

To circumvent the FIFO paradigm, the proposed hardware architecture activates read operations "N" times, depending on the needs of the user. The iteration of the read operation is determined by the kernel size. If the kernel is 3 by 3, for example, the values from three adjacent positions are read. This will determine whether there is sufficient data for IP core architecture. The memory hardware location from which we will read must be specified by the user. That is, the read output is extracted from the pre-processing block hardware as a concatenation of three pixel data values highlighted by the read pointer. Finally, data was accessed from all three locations at the same time.

4.3 Additional user choice

Image padding adds new pixels around the edges of an image. When advanced filtering methods are used, the border serves as a boundary or provides space for annotations. The three different research findings are used as user-preferred option inputs.

4.4 Duplicate mode

There are now two rows and two columns. The pixel values from the first row and the last row are copied for the new row that comes before the first row and after the last row, as shown in **Figure 2**. The first and last columns of the original image's values are also copied to the new columns. After duplication, the image would be 514 by 514 due to the cascading of two additional rows and two additional columns.

4.5 Zero padding

Figure 3 shows the addition of additional rows with all pixel values set to zero. Columns with a pixel value of 0 are also introduced.

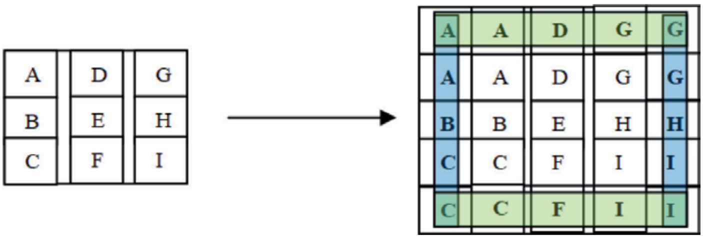


Figure 2.
Duplicate mode: original image versus with padding extra row and column.

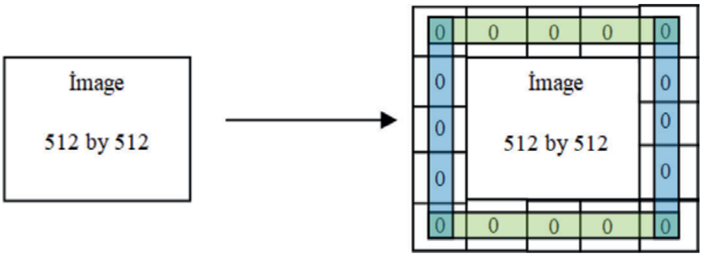


Figure 3.
Zero padding modes: original image versus with padding extra row and column filled with zero.

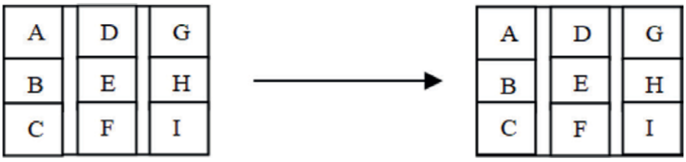


Figure 4.
Non-duplicate mode: remains same as original image (no additional row and column).

4.6 Non-duplicate mode

In this case model, function with existing/ignore the boundaries for IP core operations. Ignore the edge pixel's value and compute for those pixels that have all of their neighbors highlighted in **Figure 4**.

5. Experimental analysis of reconfigurable platform

Hardware layout strategies like parallelism and pipelining are viable on an FPGA, however now no longer in dedicated DSP modules. The use of reconfigurable hardware to put into effect photo processing algorithms lowers time-to-marketplace costs, permits speedy prototyping of complex algorithms, and makes debugging and verification easier. FPGAs are for that reason a terrific preference for real-time photo processing algorithms.

The preprocessing hardware design during this analysis paper is constructed victimization Verilog coding. The verification of practicality is performed on reconfigurable hardware using the style flowchart shown in **Figure 5**.

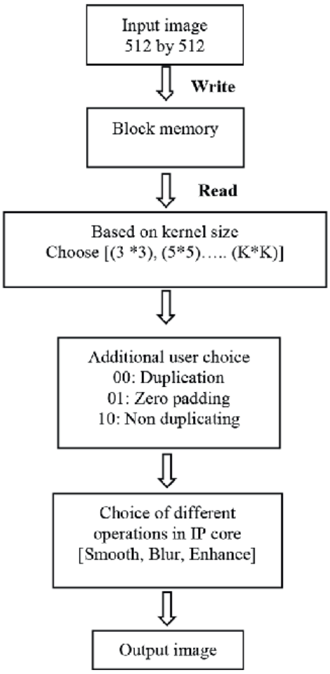


Figure 5.
Design flow for preprocessing architecture of the work.

The external view included in the 5 by 5 kernel shown in **Figure 6** is this. In this case, the output is indicated as 40 bits. That is 5 times the 8-bit pixel value depicted in the above figure in one clock cycle. It has complete control over all external input and output pins.

In digital circuit design shown in **Figure 7**, the register-transfer level (RTL) is a design abstraction for modeling a synchronous digital circuit in terms of the flow of digital signals (data) between hardware registers and the logical operations performed on those signals (**Figure 7**).

Figure 8 illustrates that FPGAs can significantly speed up image processing applications, compared to software versions. Then, by employing our method, we will be able to reduce the number of operations in a real-time application, allowing us to include more complex algorithms.

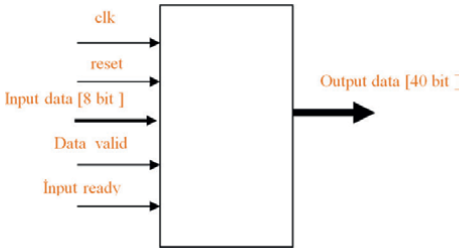


Figure 6.
External view for preprocessing architecture of the work.

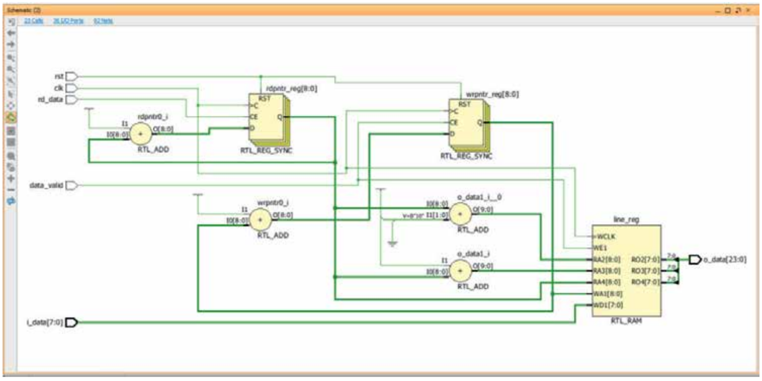
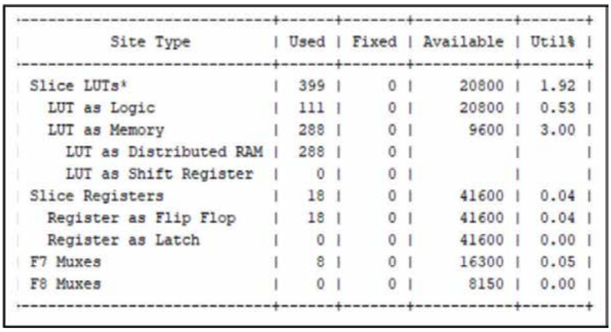


Figure 7.
RTL design for preprocessing architecture of the work.



```
+----------------------------+------+-------+-----------+-------+
|          Site Type         | Used | Fixed | Available | Util% |
+----------------------------+------+-------+-----------+-------+
| Slice LUTs*                |  399 |     0 |     20800 |  1.92 |
|   LUT as Logic             |  111 |     0 |     20800 |  0.53 |
|   LUT as Memory            |  288 |     0 |      9600 |  3.00 |
|     LUT as Distributed RAM |  288 |     0 |           |       |
|     LUT as Shift Register  |    0 |     0 |           |       |
| Slice Registers            |   18 |     0 |     41600 |  0.04 |
|   Register as Flip Flop    |   18 |     0 |     41600 |  0.04 |
|   Register as Latch        |    0 |     0 |     41600 |  0.00 |
| F7 Muxes                   |    8 |     0 |     16300 |  0.05 |
| F8 Muxes                   |    0 |     0 |      8150 |  0.00 |
+----------------------------+------+-------+-----------+-------+
```

Figure 8.
Utilization report for preprocessing architecture of the work.

6. Results and discussion

2D MAC Convolution is a general-purpose image filter effect that adds weighted values of all the pixels around it to determine the value of the center pixel. The product of the 512 * 512 image matrix convolution multiplied by [(3 * 3), (5 * 5) ..... (K * K)]. is a newly modified, filtered image. Calculate the product of the pixels that overlap each other and their sum in each case and the result will be the value of the output pixel at that particular location.

Input pixels are written into memory and accessed using a read operation in **Figure 9**; the results shown are for a 3 by 3 matrix kernel. To put it another way, it is enabling the sequential read operation three times in a row.

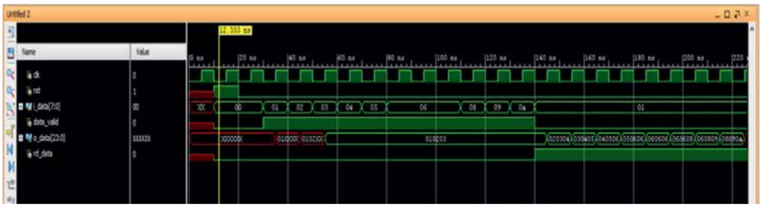


Figure 9.
Simulation results for preprocessing architecture of the work.

7. Conclusions

In this study, the preprocessing technique facilitates IP core access to data and improves image quality by using different image processing techniques. Analysis and verification of the results are carried out using a standard reconfigurable platform (Zynq board), as well as an evaluation of the consistency of hardware usage (area). The purpose of this study was to concentrate on selecting the ideal memory operations, including multiple reads and writes, and to take into consideration the main user input, the core size. Not only does preprocessing reduce memory access times, but it also improves performance. As a result of this procedure, the data collected may be useful for advancing the research. It is a reconfigurable platform that can be deployed in a variety of ways.

Acknowledgements

This research is being carried out at the Research Center of Electronics and Communications at PESIT Bangalore South Campus (PES University), Bangalore, Karnataka, India. Which is Established in 1988.

Notes/thanks/other declarations

I would like thank my guide Dr. Subhash kulkarni for his contionous support and encouragement at each every stage of research.

Author details

G.N. Chiranjeevi* and Subhash Kulkarni
PES University, Bangalore, Karnataka, India

*Address all correspondence to: chiranjeevign@pes.edu

References

[1] Smith TF, Waterman MS. Identification of common molecular subsequence's. Journal of Molecular Biology. 1981;**147**:195-197

[2] Nikhil R. Blue spec System Verilog: Efficient, correct RTL from high level specifications. In: Proceedings of the Second ACM and IEEE International Conference on Formal Methods and Models for Co-Design (MEMOCODE'04). CA, USA, San Diego. 23-25 June 2004. pp. 69-70

[3] Shoup R. Parameterized Convolution Filtering in a Field Programmable Gate Array Interval. Technical Report. Palo Alto, California; 1993

[4] Neoh HS, Hazanchuk A. Adaptive edge detection for real-time video processing using FPGAs. In: 2004 GSPx Conference. 2004

[5] Wang J, Zhong S, Yan L, Cao Z. An embedded system-on-chip architecture for real-time visual detection and matching. IEEE Transactions on Circuits and Systems for Video Technology. 2014;**24**:525-538

[6] Mondal P, Biswal PK, Banerjee S. FPGA based accelerated 3D affine transform for real-time image processing applications. Computers and Electrical Engineering. 2016

[7] Kadric E, Lakata D, Dehon A. Impact of parallelism and memory architecture on FPGA communication energy. ACM Transactions on Reconfigurable Technology and Systems. 2016;**9**:1-23

[8] Pezzarossa L, Kristensen AT, Schoeberl M, Sparsø J. Using dynamic partial reconfiguration of FPGAs in real-time systems. Microprocessors and Microsystems. 2018;**61**:198-206

[9] Stephen D. Brown RJ, Francis J, Rose ZG Vranesic. Field Programmable Gate Arrays. 1992

[10] Hirai S, Zakouji M, Tsuboi T. Implementing image processing algorithms on FPGA-based realtime vision system. In: Proceedings in 11th Synthesis and System Integration of Mixed Information Technologies (SASIMI 2003). Hiroshima; Apr 2003. pp. 378-385

[11] Torres-Huitzil C, Nuño-Maganda MA. Area time efficient implementation of local adaptive image thresholding in reconfigurable hardware. ACM SIGARCH Computer Architecture News. 2014;**42**:33-38

[12] Sungheetha A, Sharma R. A novel CapsNet based image reconstruction and regression analysis. Journal of Innovative Image Processing (JIIP). 2020;**2**(03):156-164

[13] Navabi Z. Digital Design and Implementation with Field Programmable Devices. Kluwer Academic Publishers; 2011

[14] Lysaght P, Blodget B, Mason J, Young J, Bridgford B. Invited paper: Enhanced architectures, design methodologies and CAD tools for dynamic reconfiguration of Xilinx FPGAS. In: Proceedings of the International Conference on Field Programmable Logic and Applications (FPL '06). pp. 1-6

[15] Chiranjeevi GN, Kulkarni S. Pipeline architecture for N=K*2L bit modular ALU: Case study between current generation computing and vedic computing. In: 6th International Conference for Convergence in

Technology (I2CT). 2021. pp. 1-4. DOI: 10.1109/I2CT51068.2021.9417917

[16] Durgakeri BS, Chiranjeevi GN. Implementing image processing algorithms using Xilinx system generator with real time constraints. In: 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). 2019. pp. 230-234. DOI: 10.1109/RTEICT46194.2019.9016962

[17] Ravi S, Abdul Rahim B, Fahimuddin Shaik. FPGA based design and implementation of image edge detection using Xilinx system generator. International Journal of Engineering Trends and Technology (IJETT). Oct 2013:4(10):4657-4660

[18] Raut P, Gokhale AV. FPGA implementation for image processing algorithms using Xilinx system generator. IOSR Journal of VLSI and Signal Processing; **2**(4):26-36

[19] Chaitanya Deepthi V, Surya Prasad P. Medical image fusion using Xilinx system generator in FPGA. International Journal of VLSI System Design and Communication Systems. Oct 2016;**4**(10):0990-0993

[20] Chiranjeevi GN, Kulkarni S. Validation of the FPGA-based image processing techniques using the efficient tool like Xilinx device generators. International Journal of Emerging Trends in Engineering Research. Apr 2021;**9**(4)

[21] Li C, Rui H, Zhaohua D, Chris GJ, Dimitris NMand John CG. A level set method for image segmentation in the presence of intensity in homogeneities with application to MRI. IEEE Transactions on Image Processing. 2011;**20**:2007-2016

[22] Ankita B, Kalyani M. Fuzzy-based artifcial bee colony optimization for gray image segmentation. Signal, Image and Video Processing. 2016;**10**:1089-1096

[23] Qingyi L, Mingyan J, Peirui B, Guang Y. A novel level set model with automated initialization and controlling parameters for medical image segmentation. Computerized Medical Imaging and Graphics. 2016;**48**:21-29

[24] Farid MS, Lucenteforte M, Grangetto M. DOST: A distributed object segmentation tool. Multimedia Tools and Applications. 2018;**77**:20839-20862

[25] Qureshi MN, Ahamad MV. An improved method for image segmentation using K-means clustering with neutrosophic logic. Procedia Computer Science. 2018;**132**:534-540

Edited by Morteza SaberiKamarposhti
and Mahdi Sahlabadi

Ontology is a formal characteristic and an entity of perception and concept. The characteristic of an ontology is the formal description of a set of terms and the relationships between them, which is obtained with a special language and in a file that can be understood by computers. Ontologies could be employed in several domains, such as global semantic networks, search engines, electronic commerce, natural language processing, knowledge engineering, information extraction and retrieval, multi-agent systems, qualitative modeling of physical systems, database design, information systems, and geographic and digital libraries. This book presents the latest advances and new visions in the field of ontology in information science.

Published in London, UK

IntechOpen

ISBN 978-1-80356-919-2

9 781803 569192