STANDARDIZATION OF SCIENTIFIC EXPERIMENTAL DATA REPRESENTATION THROUGH ONTOLOGY-BASED METADATA SCHEMA

NUR ADILA BINTI AZRAM

IPPH 2020 1

STANDARDIZATION OF SCIENTIFIC EXPERIMENTAL DATA REPRESENTATION THROUGH ONTOLOGY-BASED METADATA SCHEMA

By

NUR ADILA BINTI AZRAM

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy

June 2020

DEDICATED

MY PARENTS, MY FAMILY and MY FRIENDS with love

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

STANDARDIZATION OF SCIENTIFIC EXPERIMENTAL DATA REPRESENTATION THROUGH ONTOLOGY-BASED METADATA SCHEMA

By

NUR ADILA BINTI AZRAM

June 2020

Chairman : Associate Professor Rodziah Atan, PhD
Faculty : Halal Products Research Institute

Halal is a wide area that involved multidisciplinary domains such as biotechnology and medical science in which data and information come from various sources such as laboratory instruments and machines. *Halal* is defined as the status of certain products that do not contain unpermitted ingredients. The *halal* determination for various samples and ingredients were done using various laboratory instrument, at which each instrument has a different structure and format of data. These make it difficult for managing and integrating the data for analysis. Research areas involved with data management and integration need to explore data standardization as it helps in bringing data into a common format. Hence, it would help in collaborative research as well as sharing of data and information. The problem addressed in this study is, researchers in the determination of *Halal* components of products require data standardization as it is hard in managing and analyzing scientific experimental data from multiple laboratory instruments that have different structures and formats of data.

The objective of this research was to standardize scientific experimental data from Halal Institute laboratory instruments. To accomplish such goals, an ontology schema model was proposed to standardize and gives a controlled vocabulary of the scientific experimental data from Halal Institute laboratory instruments. A metadata representation structure, based on the proposed ontology schema, was also developed to give a standard structure to the scientific experimental data representation as well as simplified data that enables data retrieval and display.

Two types of evaluation were conducted in this study which was; ontology schema evaluation and metadata representation structure evaluation. Both evaluations were done using the data files from instruments in the laboratory for raw and processed materials and liquid analysis, namely Gas Chromatography-Mass Spectrometry (GC-

MS) and High Performance Liquid Chromatography (HPLC) instruments. The proposed ontology schema model was evaluated and validated based on completeness and correctness analysis measures. It was to ensure that the proposed ontology schema model was designed completely and correctly based on the grouped data and information from the laboratory instruments.

Based on the ontology schema model evaluation, the completeness percentage of the ontology schema model was 100%, conform to all of the grouped data Sample Info, Result Info, Experimental Setup Info, and Graph Info from the laboratory instruments. For the correctness percentage, the result of the ontology schema model correctly conforming to the Sample Info data of GC-MS and HPLC instruments which were 50% and 43% respectively. The correctness percentage conforms to the Result Info and Graph Info data of both instruments were 100%. For the correctness percentage conformed to the Experimental Setup Info data of GC-MS and HPLC instruments, was 96% and 86% respectively. These figures indicate that the average recall percentage of the IEDOS correctly conforms to all of the grouped data was 84%. Overall, the results gained were satisfactory although the results of the correctness percentage conform to Sample Info data was slightly lower because of data selection factors. Metadata representation structure evaluation and validation consists of precision and recall analysis to measure the accuracy of metadata extraction from the laboratory instruments data files. The precision percentages were 90% and recall were 100% for both GC-MS and HPLC instruments data files. The results gained shows the appropriate applicability of the proposed ontology-based metadata, in giving a standardized structure for the scientific experimental data for these instruments. This could positively facilitate the analysis of the scientific experimental data by giving the same structure of data to be compared and evaluated.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PEMPIAWAIAN REPRESENTASI DATA EKSPERIMENTAL SAINTIFIK MELALUI SKEMA METADATA BERASASKAN ONTOLOGI

Oleh

NUR ADILA BINTI AZRAM

Jun 2020

Pengerusi : Professor Madya Rodziah Atan, PhD
Fakulti : Institut Penyelidikan Produk Halal

Halal adalah bidang yang luas yang melibatkan domain mulitdisiplin seperti bioteknologi dan sains perubatan dimana data dan maklumat berasal dari pelbagai instrumen makmal dan mesin. Halal didefinisikan sebagai status produk tertentu yang tidak mengandungi bahan yang tidak dibenarkan. Penentuan halal untuk pelbagai sampel dan bahan dilakukan dengan menggunakan pelbagai instrumen makmal dengan setiap instrumen mempunyai struktur dan format data yang berbeza. Ini menyukarkan pengurusan dan pengintegrasian data untuk analisis. Bidang penyelidikan yang terlibat dengan pengurusan dan persepaduan maklumat perlu meneroka pempiawaian data kerana ia membantu membawa data ke dalam format yang sama. Oleh itu, ia akan membantu dalam kerjasama penyelidikan serta perkongsian data dan maklumat. Masalah yang ditangani dalam kajian ini ialah, penyelidik dalam penentuan komponen produk Halal memerlukan pempiawaian data kerana sukar dalam mengurus dan menganalisis data eksperimen saintifik dari pelbagai instrumen makmal yang mempunyai struktur dan format data yang berbeza.

Objektif penyelidikan ini adalah untuk mempiawai data saintifik eksperimen dari instrumen makmal institut Halal. Untuk mencapai matlamat tersebut, satu model skema ontologi telah ditawarkan untuk mempiawai dan memberi kawalan kepada istilah bagi data saintifik eksperimen dari instrumen makmal Halal. Satu struktur perwakilan metadata berdasarkan skema ontologi turut ditawarkan oleh penyelidikan ini untuk memberikan struktur yang piawai kepada pewakilan data saintifik eksperimen serta memudahkan data untuk membolehkan data diambil semula dan dipapar.

Dua jenis penilaian telah dilakukan dalam kajian ini iaitu penilaian skema ontologi dan penilaian struktur perwakilan metadata. Kedua-dua penilaian dilakukan mengunakan fail-fail data daripada instrumen makmal yang digunakan untuk analisis

bahan-bahan yang mentah dan sudah diproses serta bahan cecair iaitu instrumen makmal spektrometri jisim kromatografi gas (GC-MS) dan kromatografi cecair berprestasi tinggi (HPLC). Model skema ontologi yang ditawarkan dinilai dan disahkan berdasarkan analisis kelengkapan dan ketepatan. Ia adalah untuk memastikan model skema ontologi yang ditawarkan dibina dengan lengkap dan tepat berdasarkan data dan maklumat yang dikumpulkan dari instrumen makmal.

Berdasarkan penilaian model skema ontologi, peratusan kelengkapan model skema ontologi adalah 100% mematuhi semua kumpulan data maklumat sampel, maklumat hasil, maklumat persediaan eksperimen dan maklumat graf instrumen makmal. Untuk peratusan ketepatan, keputusan model skema ontologi tepat mematuhi data maklumat sampel untuk instrumen GC-MS dan HPLC masing-masing adalah 50% dan 43%. Peratusan ketepatan mematuhi data kedua-dua maklumat hasil dan maklumat graf adalah 100%. Untuk peratusan ketepatan mematuhi data maklumat persediaan eksperimen bagi instrumen makmal GC-MS dan HPLC, masing-masing adalah 96% dan 86%. Ini menunjukkan purata ketepatan model ontologi skema mematuhi semua kumpulan data adalah 84%. Keseluruhannya, keputusan yang diperolehi adalah memuaskan walaupun keputusan bagi peratusan ketepatan mematuhi data maklumat sampel agak rendah disebabkan faktor pemilihan data. Penilaian dan pengesahan struktur perwakilan metadata merangkumi analisis kepersisan dan perolehan kembali yang dilakukan untuk mengukur ketepatan metadata yang diekstrak daripada fail-fail data instrumen makmal. Keputusan kepersisan adalah 90% dan keputusan perolehan kembali adalah 100% bagi kedua-dua fail-fail data instrumen makmal GC-MS dan HPLC. Keputusan yang diperolehi menunjukkan kesesuaian keterterapan struktur yang ditawarkan dalam memberi struktur metadata yang piawai untuk data saintifik experimen daripada peralatan instrumen makmal. Ini dapat memudahkan analisis data eksperimen saintifik dengan memberikan struktur data yang sama untuk dibandingkan dan dinilai.

ACKNOWLEDGEMENTS

In the name of Allah, Most Gracious, Most Merciful

First and foremost, Alhamdulillah and thanks to Allah S.W.T for the strength, patience, courage, and determination given to me in completing this research.

I want to offer my genuine thanks and profound gratitude to my supervisor, Assoc. Prof. Dr. Rodziah Atan for her supervision, advice, recommendations, endless persistence, thoughtfulness, and eagerness to help me through the entire course of this research. Without her guidance and support, it would not be possible for me to complete my research. I have taken in a great deal of helpful information from her all through this research. I would like to also express my gratitutde to my co-supervisor, Assoc. Prof. Dr. Mohd Nasir Mohd Desa and Prof. Dr. Shuhaimi Mustafa for their thoughtfulness, helps, supports, and recommendations for this research to be finished properly.

My greatest motivation and inspiration, my mother, Taraiani Salim, my husband, Mohd Hafiz Mohd Yusoff, my three children, Airina Aireen, Azrizal Azri and Aimira Aimi, my sister, Nur Alisa Azram as well as my family in laws for their persistent love, support, and overpowering consolation to stay by my side to endure this empowering challenge.

My highest form appreciation also goes to all of the people who were willing to provide assistance and guide me through this journey. Last but not least, I would like to extend my gratitude to the Director and staff of Halal Products Research Institute for their kind cooperation in providing all necessary facilities required throughout this study.

This thesis was submitted to the Senate of the Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Rodziah Atan, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Mohd Nasir Mohd Desa, PhD
Associate Professor
Faculty of Medicine and Health Sciences
Universiti Putra Malaysia
(Member)

Shuhaimi Mustafa, PhD
Professor
Faculty of Biotechnology and Biomolecular Sciences
University Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 10 September 2020

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations, and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software

Signature: ____________________________ Date: ___________________

Name and Matric No.: Nur Adila binti Azram (GS43908)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- Supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) were adhered to.

Signature: ______________________

Name of Chairman of Supervisory Committee: Associate Professor Dr. Rodziah Atan

Signature: ______________________

Name of Member of Supervisory Committee: Associate Professor Dr. Mohd Nasir Mohd Desa

Signature: ______________________

Name of Member of Supervisory Committee: Professor Dr. Shuhaimi Mustafa

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AKHME	Adaptive Hypermedia Knowledge Management E-learning
apt	Advanced Package Tool
AVM	Astronomy Visualization Metadata
BFO	Basic Formal Ontology
BIO	Building Information Modeling
CDEs	Common Data Elements
CLI	Common Line Interface
DC	Dublin Core
DDI	Data Documentation Initiative
DF	Data File
DTO	Drug Target Ontology
DwC	Darwin Core
EHR	Electronic Health Record
EML	Ecological Metadata Language
EMR	Electronic Medical Record
EXPO	Common ontology of scientific experiments
FSA	Federal Student Aid
GC-MS	Gas Chromatography-Mass Spectrometry
GEMMS	Generic and Extensible Metadata Management System
HPLC	High-Performance Liquid Chromatography
HTML	Hypertext Markup Language
IEDMR	Instruments Experimental Data Metadata Representation
IEDOS	Instruments Experimental Data Ontology Schema
IMS	IP Multimedia Subsystem
KM	Knowledge Management
LAMBDA	Language for Metadata-Based Applications
LBS	Location-based Service
LIO	LIFE Investigation Ontology
MARC	Machine-Readable Cataloging
MATESC	Metadata-Analytic Text Extractor and Section Classifier
ME	Metadata Element
MET	Metadata Extraction Tool
MODS	Metadata Object Description Schema
NCEM	National Center for Electron Microscopy
NDBs	Standardize Nutrient Databases
OBCS	Ontology of Biological and Clinical Statistics
OBI	Ontology of Biomedical Investigations
OBO	Open Biological and Biomedical Ontologies
OCR	Optical Character Recognition
OLAC	Open Language Archives Community
OWL	Ontology Web Language
PBCore	Public Broadcasting Metadata Dictionary
RDF	Resource Description Framework
SDSS	Spatial Decision Support System
SIO	Semanticscience Integrated Ontology
SOS	Securelogy-based Self-diagnosis

TMO	Translational Medicine Ontology
TOVE	Toronto Virtual Enterprise
UAO	University Activity Ontology
UI	User Interface
VRA	Visual Resource Association
XML	Extensible Markup Language

CHAPTER 1

INTRODUCTION

1.1 Introduction

Laboratory instruments have been used by researchers and scientists to conduct scientific experiments in the various area of study to gain information and data about many things in their respective domains. All laboratory instruments have their own functions and purposes in processing various scientific experiments. Every experiment that has been carried out, produces results that need to be analyzed. Analysis of data from scientific experiments is a crucial task for researchers and scientists as it would give answers to what they want to know regarding their experiments. The incorrect analysis will negatively impact the results of the experiments they have worked on.

Analysis of scientific experiments data can sometimes be difficult to be done, as it may involve the usage of more than one laboratory instrument in conducting the experiment(s). This would involve different formats, structures, types, etc which might complicate the analysis process. There is a need to have a constant structures or formats of data that can ease the analysis process.

Thus, the main concern in this research is towards data standardization. It is defined as the process of changing data from heterogeneous sources and systems into a uniform format. It facilitates removing the incoherence in data attributes or properties which makes it easy for the management and interoperability of data. It also an essential element of advancing a research area (Brooksbank & Quackenbush, 2006). Data standardization intended at defining what information can and should be collected, determining a way to representing this information, and determining how to translate it for further transmission (Data Standardization in Healthcare: How to Adopt Data, 2017).

Standardization of data has been essential to many areas of studies that involved with a multidiscipline domain such as medical science and food technology. Scientific experimental data are generated through calculation, test process, pre-experimental design, or experimental design (Experimental Data, 2019). With multidisciplinary areas involved, scientific experimental data are stored in various platforms and human involvement in the creation of a data analysis chain is required. Collection, analysis, management, and sharing of experimental data are difficult due to informational infrastructure (Wang, Pearson, Liu, Azar, & Madlmayr, 2006) as large data volumes were daily generated with single data files in the range of several megabytes per data set produced (Potthoff, et al., 2019) at an exceptional rate from heterogeneous sources nowadays (Oussous, Benjelloun, Lahcen, & Belfkih, 2018).

There are methods and approaches used for data standardization depending on the domains. Common data elements are one of the methods that are currently used in the clinical research domain to standardize essential data. It helps to improve data quality, enhance data analysis, and support decision making. It also gives opportunities for assessment and grouping of data from multiple studies & electronic health records. Crowdsourcing mapping has also been used as one of the data standardization methods within the medical domain. It is used to standardize terminologies in medical data for medical data analysis.

Another method for data standardization is using an ontology, in which it can control the vocabularies of data terms used. It can also represent the knowledge of the standardized data into concepts within a domain and their relations (Tagliaferri, et al., 2018). It can be used by a variety of domains to represent the knowledge of the domains. Ontology helps in assigning semantic meaning to standardized data so that users would have a specific standard for a particular domain.

Ontology allows the representation of knowledge in a clear and comprehensive way due to their explicitness (Durán Muñoz & Bautista Zambrana, 2013). It was designed through the identification of the concepts and relations of a domain to represent the knowledge of the domain. It is necessary to design ontology when users need to understand data and the structure of information for other purposes such as enabling the reuse and analysis of the domain knowledge. This research will use ontology for standardization of scientific experimental data representation from laboratory instruments which would help to represent the knowledge regarding scientific experimental data as well as giving a standardized and consistent term to the scientific experimental data.

On the other hand, metadata is one of the methods that gets frequently associated to ontology in representing the semantic information about a particular data. It is often used for data management, data integration and data standardization. It is used to specify the components, scope, management, quality, the data owner as well as other relevant elements or data sets information, which is an essential element of data discovery, integration, management, exchange, and sharing (Ying & Gengda, 2004). Metadata can be described as a conceptual medium that links the contextual divide among disparate data sources (Lee, 2003).

Through the utilization of metadata, a standard structure to the data representation facilitating the users in the management and analysis of data, can be done through this research. Metadata ensures that data can be uniquely defined and correctly described to support future recovery and reuse for data management and integration. It will provide a shared interpretation of the context or semantics of the data to ensure users are using the data correctly. It also helps in giving consistency in data definitions and relations as well as in easing data representation for further analysis.

Halal industry is defined as the productions of goods and services complies to halal regulations. In the *Halal* industry, there are many information and data from different sources such as laboratory experiments, policy, and management. End-users and consumers are facing difficulties in obtaining verified information because of data stored in different sources. Additionally, researchers or scientists encounter problems in collecting, handling, analyzing, and integrating data of scientific experiments from various sources, such as laboratory instruments and machines. *Halal* laboratory is defined as facilities that are used to perform scientific or technological research, experiments, and measurement regarding *halal* products. As for *Halal* laboratory instruments, they are defined as devices that are used in the *halal* laboratory to help in performing scientific/technological research, experiments, and measurement regarding *halal* products.

Halal industry involved multidisciplinary areas such as food biotechnology and microbiology. The data or information stored may apply different structures and models. Furthermore, data can be in heterogeneous formats such as structured data, raw data, scientific data, spreadsheets, PDF files, and many others (Wang, Vergara-Niedermayr, & Liu, 2014). Thus, it is a mix of structured data and files and with different structures, models, and heterogeneous formats, will create issues on efficiency in managing and analyzing data as well as verification of the data or information.

1.2 Problem Statement

Scientific experiments involve various laboratory instruments to produce results that are used in scientific experimental data analysis. The flow of the analysis for scientific experimental data from laboratory instruments may involve numbers of different analytical instruments. Under this circumstances, human intervention is required in the experiment process and analysing the results. These interference may cause human errors within the critical analysis, affecting the end results of experiments. Moreover, these different instruments also produce different structures and formats of generated data complicates the process of analysing data.

Currently, researchers that are conducting machine-based or instrument-based experiments and analysis, are facing difficulties in managing and analyzing data due to information that is stored in isolation and different sources. Scientific experimental data often differ in terms of types, structure, requirements, methods, and processing which is shown in how different users search for and use data (Qin, Ball, & Greenberg, 2012). These make it more challenging to manage and analyze data as researchers need to consider various things regarding the format, types, or structures of the scientific experimental data involved.

In the case which researchers need to gather and analyze data from two or more different sources, human intervention may be required as the sources may apply different formats and structures of data representation. Data stored in various file formats will make data management and interchange prohibitively laborious (Ingargiola, Laurence, Boutelle, Weiss, & Michalet, 2016). These might as well cause inaccuracies in analyzing the data which results in repeating the experiments and directly increase the experimental cost and relevantly time-consuming. Researchers need to have a full understanding of data gathered and the assume relationships to conduct the analysis for results and conclusions.

1.3 Objectives

This research aimed at standardizing scientific experiment data representation from *Halal* laboratory instruments by modeling an ontology schema; which provides a standardized and controlled vocabulary of the terms for instruments experimental data. This research will also build a metadata representation structure based on the ontology schema as a standard structure to represent scientific experimental data from *Halal* laboratory instruments. This structure is expected to facilitate researchers in managing and analyzing their experiment(s) data.

The objectives of this research are as follows:

1. To model an ontology schema to standardize the scientific experimental data and gives a controlled set of vocabulary data terms from *Halal* laboratory instruments.

2. To build a metadata representation structure as a standard structure to represent the scientific experimental data from *Halal* laboratory instruments.

3. To evaluate the metadata extraction of scientific experimental data through precision and recall analysis for accuracy validation.

1.4 Scope and significance of the study

In this research, the standardization of scientific experimental data representation, which focuses on the restructuring of scientific experimental data representation for analysis, from Halal laboratory instruments has been made. *Halal* has been an emerging sector and is steadily reaching out among scholars and practitioners (Haleem, Khan, Khan, & Jami, 2019). It involved numerous data and information obtained from different sources such as laboratory experiments that were important in giving information on samples related to *Halal*. Management and analysis of the data and information from *Halal* laboratory are difficult to be processed due to different instruments producing different data (Rudd, 2017). Additionally, there are also lack of uniform data standards, consistent description format, and presentation methods

(Liang, et al., 2018). As there is lack of research on the standardization of scientific experimental data within the *Halal* scope, it would be a great opportunity to expands the research to caters problems of management and analysis of data from *Halal* laboratory instruments. Hence, these are the main concerns on selecting *Halal* as the preferred industry for this research.

Two instruments were selected in designing the ontology schema of instruments experimental data. The instruments were Gas Chromatography-Mass Spectrometry (GC-MS) and High-Performance Liquid Chromatography (HPLC). GC-MS separates chemical mixtures (the GC component) and identifies the components at a molecular level (the MS component). It is used in determining the alcohol content from samples such as fermented foods and carbonated drinks (which in the case of determining the samples *Halal* status, the alcohol content needs to be below 1%) as well as analyzing a variety of oils such as essential oils and perfumes. HPLC is an analytical technique to separate, identify, and quantify components in a mixture. It is used for amino acid and protein profiling from samples such as collagen, gelatine, and capsule.

This research has pointed out the usage of both GC-MS and HPLC since they are popular instruments and frequently required for sample testing. They have a similar purpose and are often used together in experiments, either concurrently or sequentially to compare similar sample testing or experiments. These reasons highlight the reason why GC-MS and HPLC were the equipment studied in this research instead of other instruments such as FTIR and GC-TOF that are also easily available in the laboratory. Both instruments also function under the same basic principles of compound separation, identification, and quantification techniques. For the developed metadata extraction tool, the extraction currently uses text file data type.

This study is significant to researchers' that deal with scientific experimental data by providing data analysis with a standard structure view of data through the proposed ontology-based metadata. This study can be a valuable source of information for domains that works with the standardization of scientific experimental data from laboratory instruments. The ontology-based metadata can be used in other domains, which uses similar laboratory instruments as these instruments were not specifically for *Halal* industries and are regularly applicable in other domains.

1.5 Thesis Outline

This thesis is outlined into six chapters, including this chapter which provides a background of data standardization, metadata, and ontology. This chapter also presented the problem statement and list of objectives, scope, and significance of this research.

Chapter two consists of literature reviews on the related work of this research. It begins with a description of data and laboratory instruments as scientific data sources as well as data standardization. It also describes and reviews ontology and metadata in detail for further apprehension. It also illustrates ontology validation, metadata validation and laboratory instruments analysis.

Chapter three is an overall description of the research methodology. It describes the flow of activities devised for this research. It also describes the instruments' data examples and the research framework. It explained the evaluation and validation steps for the proposed ontology schema as well as the analysis method for the ontology evaluation. The evaluation and validation steps for the proposed metadata structure as well as an analysis method for the metadata extraction, were also discussed in the chapter.

Chapter four discusses the research designs of the research. It begins with the description of the proposed ontology schema and details of ontology evaluation. Followed by the discussion on the metadata representation structure based on the proposed ontology. The metadata extraction tool development and details on metadata representation structure evaluation were also depicted in depth.

Chapter five presents the analysis of the obtained results by the ontology schema evaluation and metadata extraction evaluation. This chapter also discusses the result of ontology schema validation using correctness and completeness analysis as well as accuracy validation results for the metadata extraction using precision and recall analysis.

Chapter six, elaborates on reflected conclusion and contributions of the study. The list of appropriate recommendations for future research works were also provided.

REFERENCES

Abdalla, A., Afify, A., Hasaan, I., & Mohamed, A. (2018). Studying the Effect of Household-Type Treatment and Processing on the Residues of Ethion and Profenofos Pesticides and on the Contents of Capsaicinoids in Green Chili Pepper Using GC-MS/MS and HPLC. *Food analytical methods*, 382-393.

Akimoto, R., & Kameyama, W. (2006). A Study on Language Design of Creating Compound Metadata Schama Method Based on Museum Information. *Proceedings of the 5th WSEAS International Conference on E-ACTIVITIES*, (pp. 121-126). Venice.

Alexander, P. (2019, March 20). *The Importance of Ontologies.* Retrieved from In The MMI Guides: Navigating the World of Marine Metadata: http://uop.whoi.edu/techdocs/presentations/MMI_Guides.pdf

Avila, C., Maia, G., Franco, W., Vidal, T., Franco, A., & Vidal, V. (2018). OntoVal: A Tool for Ontology Evaluation by Domain Specialists. *38th International Conference on Conceptual Modeling*, (pp. 143-147).

Baca, M., Gilliland, A., Gill, T., Woodley, M., & Whalen, M. (2008). *Introduction to Metadata (2nd Edition).* Los Angeles, USA: Getty Publications.

Blázquez, M., Fernández-López, M., García-Pinar, J., & Gomez-Perez, A. (1998). Building Ontologies at the Knowledge Level using the Ontology Design Environment. *Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'98).* Banff, Alberta, Canada.

Borst, W. (1997). *Constructive of Engineering Ontologies for Knowledge Sharing and Reuse*. Enschede: Centre of Telematics and Information Technology (CTIT).

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data Driven Ontology Evaluation. *In: International Conference on Language Resources and Evaluation (LREC 2004).* Lisbon, Portugal.

Brooksbank, C., & Quackenbush, J. (2006). Data Standards: A Call to Action. *Omics A Journal of Integrative Biology* , 94-99.

Bukhari, S., Martínez-Romero, M., O' Connor, M., Egyedi, A., Willrett, D., Graybeal, J., . . . Kleinstein , S. (2018). CEDAR OnDemand: a browser extension to generate ontology-based scientific metadata. *BMC Bioinformatics*.

Bychkov, I., Dubenskaya, J., Korosteleva, E., Kryukov, A., Mikhailov, A., Nguyen, M.-D., & Shigarov, A. (2019). Metadata Extraction from Raw Astroparticle Data of TAIGA Experiment. *Proc. 3rd Int. Workshop on Data Life Cycle in Physics*, (pp. 26-34).

Chauhan, A., Goyal, M., & Chauhan, P. (2014). GC-MS Technique and its Analytical Applications in Science and Technology. *Journal of Analytical and Bioanalytical Techniques*.

Chitkara, A., Singh, D., Gupta, A., & Varshney, G. (2020). IntelliSpect: Personal Information Search Tool. *2020 International Conference on Information Networking (ICOIN)*, (pp. 556-561).

Coyle, K., & Baker, T. (2008). *DCMI: Guidelines for Dublin Core Application Profiles*. Retrieved from http://dublincore.org/documents/2009/05/18/profile-guidelines/

Creating a metadata schema - where to start. (2018). Retrieved from International Organization for Standardization (ISO): https://committee.iso.org/files/live/sites/tc46sc11/files/documents/N800R1%20Where%20to%20start-advice%20on%20creating%20a%20metadata%20schema.pdf

da Cruz, S., Campos, M., & Mattoso, M. (2012). A foundational ontology to support scientific experiments. *CEUR Workshop Proceedings*, (pp. 144-155).

Dalal, M., Yadav, S., & Yadav, K. (2017). Role of Metadata in Knowledge Management of Multinational Organizations. *Advances in Computational Sciences and Technology*, 211-219.

Data standardization - Data transformation. (2015, September 10). Retrieved from Experian: https://www.edq.com/data-quality-management/data-standardization/

Data Standardization in Healthcare: How to Adopt Data. (2017, July 10). Retrieved from Archer - We Develop Success: https://www.archer-soft.com/en/blog/data-standardization-healthcare-how-adopt-data

Data Standardization Policies and Procedures. (2017, May 6). Retrieved from Federal Student Aid: https://studentaid.ed.gov/sa/sites/default/files/fsawg/static/gw/docs/ciolibrary/Data_Standardization_Policies_and_Procedures.pdf

Ding, L., Zhong, B., Wu, S., & Luo, H. (2016). Construction risk knowledge management in BIM using ontology and semantic web technology. *Safety Science*, 202-213.

Dublin Core. (2019, February 19). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Dublin_Core

Dumontier, M., Baker, C., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., . . . Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*.

Durán Muñoz, I., & Bautista Zambrana, M. (2013). Applying Ontologies to Terminology: Advantages and Disadvantages. *Hermes-Journal of Language and Communication in Business*, 65-77.

Euzenat, J. (2007). Semantic Precision and Recall for Ontology Alignment Evaluation. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, (pp. 348-353). Hyderabad, India.

Experimental Data. (2019, 6 January). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Experimental_data

Fernández-López, M. (1999). Overview Of Methodologies For Building Ontologies. *Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5).* Stockholm, Sweden.

Fernández-López, M., Gómez-Pérez, A., & Juristo, N. (1997). METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series.* Stanford University, EEUU.

Fernandez-Lopez, M., Gomez-Perez, A., Sierra, J., & Pazos, A. (1999). Building a Chemical Ontology Using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems* (pp. 37-46). IEEE.

Gai, K., Qiu, M., Jayaraman, S., & Tao, L. (2015). Ontology-Based Knowledge Representation for Secure Self-Diagnosis in Patient-Centered Telehealth with Cloud Systems. *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing* (pp. 98-103). New York, USA: IEEE.

Gauthier, A., Christensen, L., Hurt, R., & Wyatt, R. (2008). Virtual Astronomy Multimedia Project. *In CAP2007 Conference Proceedings*, (pp. 214-219).

Gayathri, R., & Vijayasundaram, U. (2018). Ontology based knowledge representation technique, domain modeling languages and planners for robotic path planning: A survey. *ICT Express*, 69-74.

Giri, K. (2011). Role of Ontology in Semantic Web. *DESIDOC Journal of Library & Information Technology*, 116-120.

Gomez-Perez, A. (1998). Knowledge sharing and reuse. *In J. Liebowitz, editor, TheHandbook of Applied Expert Systems*.

Goranova, M., Shishedjiev, B., & Georgieva, J. (2011). Research on Building Scientific Data Ontology. *2011 Developments in E-systems Engineering* (pp. 541-546). Dubai, United Arab Emirates: IEEE.

Gruber, T. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, 907-928.

Gruber, T. (2009). Ontology in Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.). *Springer-Verlag*.

Gruninger, M., & Fox, M. (1995). Methodology for the Design and Evaluation of Ontologies. *In Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*. Montreal.

Gueguen, G. (2010, July 2). *MODS and MADS: Current implementations and future directions*. Retrieved from LITA Blog: https://litablog.org/2010/07/mods-and-mads-current-implementations-and-future-directions/

Guenther, R. (2004). Using the Metadata Object Description Schema (MODS) for resource description: guidelines and applications. *Library Hi Tech*, 89-98.

Hamilton, D., Wheeler, D., White, C., Rees, C., Komendantov, A., Bergamino, M., & Ascoli, G. (2017). Name-calling in the hippocampus (and beyond): coming to termswith neuron types and properties. *Brain Informatics*, 1-12.

Hansen, L., Lanza, D., & Pasco, S. (2012). Developing an ontology for standardizing space systems data exchange. *2012 IEEE Aerospace Conference*. Big Sky, MT, USA: IEEE.

Hao, C., Zhao, X., & Yang, P. (2007). GC-MS and HPLC-MS analysis of bioactive pharmaceuticals and personal-care products in environmental matrices. *TrAC Trends in Analytical Chemistry*, 569-580.

Hartmann, J., Palma, R., & Gomez-Perez, A. (2010). Ontology Repositories. In *Handbook on Ontologies* (pp. 551-571). Springer.

He, P., & Aga, D. (2019). Comparison of GC-MS/MS and LC-MS/MS for the analysis of hormones and pesticides in surface waters: advantages and pitfalls. *Analytical Methods*, 1436-1448.

Heidorn, P., & Wei, Q. (2008). Automatic Metadata Extraction from Museum Specimen Labels. *In International Conference on Dublin Core and Metadata Applications*, (pp. 57-68).

Hirwade, M. (2011). A study of metadata standards. *Library Hi Tech News*, 18-25.

Hsu, L., Martin, R., McElroy, B., Kim, W., & Litwin Miller, K. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. *Geomorphology*, 180-189.

Huynh, T., & Hoang, K. (2010). GATE framework based metadata extraction from scientific papers. *In 2010 International Conference on Education and Management Technology* (pp. 188-191). IEEE.

Ingargiola, A., Laurence, T., Boutelle, R., Weiss, S., & Michalet, X. (2016). Photon-HDF5: open data format and computational tools for timestamp-based single-molecule experiments. *In´ Single Molecule Spectroscopy and Superresolution Imaging IX* (p. 971405). International Society for Optics and Photonics.

Jaswir, I., Mirghani, M., Salleh, H., Ramli, N., Octavianti, F., & Hendri, R. (2016). An Overview of the Current Analytical Methods for Halal Testing. *In´ Contemporary Issues and Development in the Global Halal Industry* (pp. 291-300). Singapore: Springer.

Jing, L., An-ronga, D., & Qi-zhi, M. (2008). THE STUDY OF INTEGRATION OF MULTI-SOURCES HETEROGENEOUS DATA BASED ON THE ONTOLOGY. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 571-576.

Kauer, J., & Whte, J. (2009). Electronic Nose. In *Reference Module in Neuroscience and Biobehavioral Psychology. Encyclopedia of Neuroscience* (pp. 871-877).

Kawtrakul, A., & Yingsaeree, C. (2005). A unified framework for automatic metadata extraction from electronic document. *In Proceedings of The International Advanced Digital Library Conference.* Nagoya, Japan.

Khankasikam, K. (2011). Metadata Extraction Using Case-based Reasoning for Heterogeneous Thai Documents. *International Journal of Computer and Electrical Engineering*.

Kiani, S., Minaei, S., & Ghasemi-Varnamkhasti, M. (2018). Instrumental approaches and innovative systems for saffron quality. *Journal of Food Engineering*, 1-10.

Kirsten, T., Kiel, A., Rühle, M., & Wagner, J. (2017). Metadata Management for Data Integrationin Medical Sciences. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)*, 175-194.

Kiweler, M., Looso, M., & Graumann, J. (2018). MARMoSET – Extracting Publication-Ready Mass Spectrometry Metadata from RAW Files. *Molecular & Cellular Proteomics*, 1700-1702.

Le Franc, Y., Bandrowski, A., Brůha, P., Papež, V., Grewe, J., Mouček, R., . . . Thomas, W. (2014). Describing neurophysiology data and metadata with OEN, the Ontology for Experimental Neurophysiology. *Frontiers in Neuroinformatics*.

Lee, C.-I., Hsia, T.-C., Hsu, H.-C., & Lin, J.-Y. (2017). Ontology-based tourism recommendation system. *2017 4th International Conference on Industrial Engineering and Applications (ICIEA)* (pp. 376-379). IEEE.

Lee, K., Lee, J., & Kwan, M.-P. (2017). Location-based service using ontology-based semantic queries: A study with a focus on indoor activities in a university context. *Computers, Environment and Urban Systems*, 41-52.

Lee, P. W. (2003). *Metadata representation and management for context mediation.* Massachusetts Institute of Technology.

Liang, H., Luo, M., Wang, R., Lu, P., Luand, W., & Lu, L. (2018). Big Data in Health Care: Applications and Challenges. *Data and Information Management*.

Lin, Y., Mehta, S., Mcginty, H., Turner, J., Vidovic, D., Forlin, M., . . . Schurer, S. (2017). Drug target ontology to classify and integrate drug discovery data. *Journal of Biomedical Semantics*.

Luciano, J., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C., . . . Dumontier, M. (2011). The Translational Medicine Ontology and Knowledge Base: Driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics*.

Lüdtke, D., Mece, S., Deshmukh, M., Bock, M., Schreiber, A., & Gerndt, A. (2012). A Framework to Model Metadata for Knowledge Management Tools. *In 4th International Conference on Knowledge Management for Space Missions (Toulouse Space Show).* Toulouse, France.

Marcos, J., & Pozo, O. (2015). Derivatization of steroids in biological samples for GC-MS and LC-MS analyses. *Bioanalysis*, 2515-2536.

Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Bürkle, T., & Ganslandt, T. (2015). Ontology-Based Data Integration between Clinical and Research Systems. *PLoS One*.

Merchant, A., & Dehghan, M. (2006). Food composition database development for between country comparison. *Nutrition Journal*.

Métral, C., Falquet, G., & Vonlanthen, M. (2007). An Ontology-based Model for Urban Planning Communication. In *Ontologies for Urban Development* (pp. 61-72). Springer.

Michener, W. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, 3-7.

Michener, W., Brunt, J., Helly, J., Kirchner, T., & Stafford, S. (1997). Nongeospatial Metadata for the Ecological Sciences . *Ecological Applications*, 330-342.

Moreno-Arribas, M., & Polo, M. (2003). CHROMATOGRAPHY | High-performance Liquid Chromatography. In *Encyclopedia of Food Sciences and Nutrition (Second Edition).*

Muse, D. (2017, June 20). *Structured Data*. Retrieved from Datamation: https://www.datamation.com/big-data/structured-data.html

Nguyen, A., Gardner, L., & Sheridan, D. (2019). Towards Ontology-Based Design Science Research for Knowledge Accumulation and Evolution. In Proceedings of the 52nd Hawaii International Conference on System Sciences, (pp. 5755-5764).

Noy, N., & McGuinness, D. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology.* Stanford Knowledge Systems Laboratory.

Ochiai, K., Nagamori, M., & Sugimoto, S. (2014). A Metadata Schema Design Model and Support System Based on an Agile Development Model. *iConference 2014 Proceedings*.

Oerlemans, A. (2011, December 22). *Content-based retrieval of visual information.* Leiden, Netherlands: Leiden Institute for Advanced Computer Sciences (LIACS), Faculty of Science, Leiden University.

Øhrstrøm, P., & Schärfe, H. (2004). A Priorean Approach to Time Ontologies. *International Conference on Conceptual Structures* (pp. 388-401). Springer.

Øhrstrøm, P., Andersen, J., & Henrik , S. (2005). What Has Happened to Ontology. *International Conference on Conceptual Structures* (pp. 425-438). Berlin, Heidelberg: Springer.

Ojokoh, B., Adewale, O., & Falaki, S. (2009). Automated document metadata extraction. *Journal of Information Science*, 563-570.

Oussous, A., Benjelloun, F.-Z., Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 431-448.

Perez-Rey, D., Maojo, V., Garcia-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martin-Sanchez, F., & Sousa, A. (2006). ONTOFUSION: ontology-based integration of genomic and clinical databases. *Computers in Biology and Medicine*, 712-730.

Pickell, D. (2018, November 16). *Structured vs Unstructured Data – What's the Difference?* Retrieved from Learning Hub: https://learn.g2crowd.com/structured-vs-unstructured-data

Pinto, H., & Martins, J. (2004). Ontologies: How can They be Built? *Knowledge and Information Systems*, 441-464.

Polavaram, S., & Ascoli, G. (2017). An ontology-based search engine for digital reconstructions of neuronal morphology. *Brain Informatics*, 123-134.

Potthoff, J., Tremouilha, P., Hodapp, P., Neumair, B., Bräse, S., & Jung, N. (2019). Procedures for systematic capture and management of analytical data in academia. *Analytica Chimica Acta: X*.

Psyllidis, A. (2015). OSMoSys: A Web Interface for Graph-Based RDF Data Visualization and Ontology Browsing. *Engineering the Web in the Big Data Era: 15th International Conference, ICWE 2015*, (pp. 679-682). Rotterdam, The Netherlands.

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. *Proceedings of the 2012 International Conference on Dublin Core and Metadata Applications*, (pp. 62-71).

Quix, C., Hai, R., & Vatov, I. (2016). Metadata Extraction and Management in Data LakesWith GEMMS. *Complex Systems Informatics and Modeling Quarterly*, 67-83.

Ram, S., & Rao, N. (2014). Metadata Description Framework for Integration of Bioinformatics Information Resources: A Case of iBIRA. *DESIDOC Journal of Library & Information Technology*, 384-392.

Ranzinger, R., Aoki-Kinoshita, K., Campbell, M., Kawano, S., Lütteke, T., Okuda, S., . . . Narimatsu, H. (2015). GlycoRDF: An ontology to standardize Glycomics data in RDF. *Bioinformatics*, 919-925.

Rautenberg, S. (2012). Modelo de conhecimento para mapeamento de instrumentos da gestão do conhecimento e de agentes computacionais da engenharia do conhecimento baseado em ontologias.

Rego, H., Moreira, T., Morgado, E., & García-Peñalvo, F. (2010). Metadata and Knowledge Management driven Web-based Learning Information System towards Web/e-Learning 3.0. *International Journal of Emerging Technologies in Learning (iJET)*, 36-44.

Riley, J. (2004, January 1). *Understanding Metadata.* Retrieved from National Information Standards Organization (NISO): https://www.niso.org/publications/understanding-metadata

Rocca-Serra, P., Salek, R., Arita, M., Correa, E., Dayalan, S., Gonzalez-Beltran, A., . . . Neumann, S. (2016). Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics*.

Rodrigo, G., Henderson, M., Weber, G., Ophus, C., Antypas, K., & Ramakrishnan, L. (2018). ScienceSearch: Enabling Search through Automatic Metadata Generation. *2018 IEEE 14th International Conference on e-Science* (pp. 93-104). IEEE.

Roe, C. (2012, June 7). *A Short History of Ontology: It's not just a Matter of Philosophy Anymore*. Retrieved from Dataversity: https://www.dataversity.net/a-short-history-of-ontology-its-not-just-a-matter-of-philosophy-anymore/

Rübel, O., Dougherty, M., Prabhat, Denes, P., Conant, D., Chang, E., & Bouchard, K. (2016). Methods for Specifying Scientific Data Standards and Modeling Relationships with Applications to Neuroscience. *Frontiers in neuroinformatics*.

Rubin, D., Noy, N., & Musen, M. (2007). Protégé: A Tool for Managing and Using Terminology in Radiology Applications. *Journal of Digital Imaging*, 34-46.

Rudd, J. (2017, May 2). *7 Data Challenges in the Life Sciences*. Retrieved from Informatics from Technology Networks: https://www.technologynetworks.com/informatics/lists/7-data-challenges-in-the-life-sciences-288265

Rushing, B., Wooten, A., Shawky, M., & Selim, M. (2016). Comparison of LC-MS and GC-MS for the Analysis of Pharmaceuticals and Personal Care Products in Surface Water and Treated Wastewaters. *Current Trends in Mass Spectrometry*, 8-14.

Samwald, M., Giménez, J., Boyce, R., Freimuth, R., Adlassnig, K.-P., & Dumontier, M. (2015). Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. *BMC medical informatics and decision making*.

Santos, W., de Almeida, E., & Meira, S. (2012). TIRT: A Traceability Information Retrieval Tool for Software Product Lines Projects. *In 2012 38th Euromicro Conference on Software Engineering and Advanced Applications* (pp. 93-100). IEEE.

Schembera, B., & Iglezakis, D. (2020). EngMeta: Metadata for Computational Engineering. *International Journal of Metadata, Semantics and Ontologies*, 26-38.

Schomburg, I., Chang, A., & Schomburg, D. (2014). Standardization in enzymology—Data integration in the world′s enzyme information system BRENDA. *Perspectives in Science*, 15-23.

Shen, H., Jiang, H., Mao, H., Pan, G., Zhou, L., & Cao, Y. (2007). Simultaneous determination of seven phthalates and four parabens in cosmetic products using HPLC-DAD and GC-MS methods. *Journal of separation science*, 48-54.

Singh, I., Kuscuoglu, M., Harkins, D., Sutton, G., Fouts, D., & Nelson, K. (2019). OMeta: an ontology-based, data-driven metadata tracking system. *BMC Bioinformatics* .

Sir, M., Bradac, Z., & Fiedler, P. (2015). Ontology versus Database. *IFAC-PapersOnLine*, 220-225.

Soldatova, L., & King, R. (2007). An Ontology of Scientific Experiments. *Journal of The Royal Society Interface*, 795-803.

Southgate, D., & Greenfield, H. (1992). Principles for the Preparation of Nutritional Data Bases and Food Composition Tables. In *International Food Data Bases and Information Exchange. Concepts, Principles and Designs* (pp. 27-48).

Studer, R., Benjamins, V., & Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & Knowledge Engineering*, 161-197.

Tagliaferri, L., Gobitti, C., Colloca, G., Boldrini, L., Farina, E., Furlan, C., . . . Mangoni, M. (2018). A new standardized data collection system for interdisciplinary thyroid cancer management: Thyroid COBRA. *European Journal of Internal Medicine*, 73-78.

Tan, H., Adlemo, A., Tarasov, V., & Johansson, M. (2017). Evaluation of an Application Ontology. *In Proceedings of the Joint Ontology Workshops 2017 Episode 3: The Tyrolean Autumn of Ontology Bozen-Bolzano.* Italy.

Tartir, S., Arpinar, I., & Sheth, A. (2010). Ontological Evaluation and Validation. In *Theory and Applications of Ontology: Computer Applications* (pp. 115-130).

Ting, K. (2011). Precision and Recall. In C. Sammut, & G. Webb, *Encyclopedia of Machine Learning.* Boston, MA: Springer.

Tkaczyk, D., Szostek, P., Dendek, P., Fedoryszak, M., & Bolikowski, Ł. (2014). CERMINE -- Automatic Extraction of Metadata and References from Scientific Literature. *11th IAPR International Workshop on Document Analysis Systems* (pp. 217-221). IEEE.

Torre, M., Aguirre, C., Anshutz, B., & Hsu, W. (2018). MATESC: Metadata-Analytic Text Extractor and Section Classifier for PDF Scientific Publications. *IC3K 2018 - Proceedings of the 10th International Joint Conference on Knowledge*, 261-267.

Uciteli, A., & Kirsten, T. (2015). Ontology-based retrieval of scientific data in LIFE. *Datenbanksysteme für Business, Technologie und Web (BTW 2015)*, (pp. 109-114).

Uschold, M. (1996). Building Ontologies: Towards a Unified Methodology. *In Proceedings of 16th Annual Conference of the British Computer Society Specialists Group on Expert Systems.*

Uzdanaviciute, V., & Butleris, R. (2011). Ontology-based Foundations for Data Integration. *In BUSTECH The First International Conference on Business Intelligence and Technology*, (pp. 34-39).

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. *International Journal of Digital Curation*.

Vasanthapriyan, S., Tian, J., Zhao, D., & Xiong, S. (2017). An Ontology-Based Knowledge Sharing Portal for Software Testing. *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (pp. 472-479). Prague, Czech Republic: IEEE.

Verma, A. (2016). An abstract framework for ontology evaluation. *2016 International Conference on Data Science and Engineering (ICDSE)* (pp. 1-6). Cochin, India: IEEE.

(2002). *VRA Core Categories Version 3.0.* Visual Resources Association.

Wang, F., Pearson, J., Liu, P., Azar, F., & Madlmayr, G. (2006). Experiment Management with Metadata-based Integration for Collaborative Scientific Research. *22nd International Conference on Data Engineering.* IEEE.

Wang, F., Vergara-Niedermayr, C., & Liu, P. (2014). Metadata based management and sharing of distributed biomedical data. *International Journal of Metadata Semantics and Ontologies*, 42-57.

West Jr, L., & Hess, T. (2002). Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial data. *Decision Support Systems*, 247-264.

While, A., Baker, A., Bloss, M., Burrows, P., Efthimiadis, E., Brooks, M., . . . Twohill, C. (2003). PB Core--the Public Broadcasting Metadata Initiative: Progress Report. *DCMI '03 Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice - metadata research & applications* . Seattle, Washington.

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

Wiser, S., Spencer, N., De Cáceres, M., & Kleikamp, M. (2011). Veg-X – An exchange standard for plot-based vegetation data. *Journal of Vegetation Science*, 598-609.

Wolff, C. (1736). *Philosophia prima sive Ontologia.* Prostat in officina libraria Rengeriana.

Wuest, T., Tinscher, R., Porzel, R., & Thoben, K.-D. (2014). Experimental Research Data Quality in Materials Science. *International Journal of Advanced Information Technology*, 1-18.

Yeumo, E., Alaux, M., Arnaud, E., Aubin, S., Baumann, U., Buche, P., . . . Quesneville, H. (2017). Developing data interoperability using standards: A wheat community use case. *F1000 Research*.

Yew, K., Hassan, M., & M. Shariff, A. (2015). Ontology evaluation — A criteria selection framework. *2015 International Symposium on Mathematical Sciences and Computing Research (iSMSC)*, (pp. 298-303).

Ying, Y., & Gengda, J. (2004). Metadata-based information organization and ontology-based knowledge organization. *Journal of Academic Libraries*, 43-47.

Yuan, E. (2017). Towards Ontology-Based Software Architecture Representations. *2017 IEEE/ACM 1st International Workshop on Establishing the Community-Wide Infrastructure for Architecture-Based Software Engineering (ECASE)* (pp. 21-27). Buenos Aires, Argentina : IEEE.

Zaman, S., Sarntivijai, S., & Abernethy, D. (2017). Use of Biomedical Ontologies for Integration of Biological Knowledge for Learning and Prediction of Adverse Drug Reactions. *Gene regulation and systems biology* , 1-7.

Zhang, Y.-F., Tian, Y., Zhou, T.-S., Araki, K., & Li, J.-S. (2016). Integrating HL7 RIM and Ontology for Unified Knowledge and Data Representation in Clinical Decision Support Systems. *Computer Methods and Programs in Biomedicine*, 94-108.

Zheng, J., Harris, M., Masci, A., Lin, Y., Hero, A., Smith, B., & He, Y. (2016). The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis. *Journal of Biomedical Semantics*.