Semantic Analysis Mapping Framework for Clinical Coding Schemes: A Design Science Research Approach

A dissertation submitted to the College of Communication and Information of Kent State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Julaine Clunis

December 2021

Dissertation written by Julaine S. Clunis

B.Sc., Northern Caribbean University, 2005

M.L.I.S., Kent State University, 2016

M.S., Kent State University, 2016

Ph.D., Kent State University, 2021

Approved by

_____

Marcia Lei Zeng, Ph.D., Chair, Doctoral Dissertation Committee

_____

Athena Salaba, Ph.D., Member, Doctoral Dissertation Committee

_____

Rebecca Meehan, Ph.D., Member, Doctoral Dissertation Committee

_____

Yi Hong, Ph.D., Member, Doctoral Dissertation Committee

_____

Mary Anthony, Ph.D., Member, Doctoral Dissertation Committee

Accepted by

_____

Miriam Matteson, Ph.D., Chair, Doctoral Studies Committee, College of Communication and Information

_____

Amy Reynolds, Ph.D., Dean, College of Communication and Information

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my family for constantly encouraging me and keeping me focused on this journey. Special thanks to my mom, Cherine Clunis, for the sacrifice, tears, and prayers that have made this possible; I have my testimony. To my other mother mentors, Carolyn Smith who started me on this path and Marva Shand McIntosh, who have kept me prayed up, and been a listening ear and a voice of experience, thank you so much. I also want to express my gratitude to Ken and Cindy Ferguson, and David and Christina Humble for being my family away from home, a cheer squad, and invaluable support system. Special thanks to Kim Batzer for more reasons than I can mention. In addition, I want to extend my heartfelt gratitude and appreciation to Jimmy for walking this tough road with me. To my advisor, Dr. Marcia Zeng, thank you for choosing me, nurturing my dreams and interests, providing me with opportunities and spending so much time and effort to enhance my learning and to ensure I do good work. You've gone above and beyond, and I salute you. Thank you to Dr. Yi Hong and Dr. Tao Hu for lending your expert knowledge to helping me through this process. I'd like to express my appreciation to the members of my dissertation committee who worked so well together and made sure I had all I needed to get things done. Finally, to all the faculty of the School of Information, particularly to Dr. Rebecca Meehan who acted as a second advisor and friend to me and whose smile, and cheerful, positive spirit has buoyed me up on many low days, thank you so much.

Glossary of Terms

Term	Definition
Administrative terminology	Coding schemes which support administrative functions such as billing, insurance reimbursement, the collection of secondary data.
Alignment	Alignment is a set of correspondences between two or more ontologies achieved through the matching process.
Annotation	Associating labels to a document and its contents to identify entities, relationships, sentiments et cetera
Binary classification	The process whereby a classifier is trained on a set of alignments to make predictions of semantic equivalence between concepts.
Cancer	A group of diseases characterized by the uncontrolled growth and spread of abnormal cells, the spread of which, if not controlled, can result in death.
Classification systems	Hierarchical and faceted arrangements of numerical or alphabetical notations to represent broad topics.
Clinical coding schemes	Structured lists of terms and their associated definitions that are intended to describe the healthcare domain categorically.
Concept	A term describing a task, function, action, strategy, reasoning process to be expressed relative to other concepts. They can be implicit or explicit, with their explicit

	definition being expressed through simulated knowledge, description logic, and concept maps.
Conceptual model	A visual representation of theoretical constructs (and variables) or system made of a composition of concepts of interest in a certain domain.
Controlled terminologies	An organized arrangement of words and phrases used to index content or retrieve content through browsing or searching.
Coordination	A characteristic of KOS that describes how the terms or concepts in the scheme are combined.
Data sharing	The ability to share the same data resource with multiple applications or users.
Description logic	A family of knowledge representation languages that are widely used in ontological modeling.
Design Science Research	A research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence
Entities	A representation of an object or thing.
F measure	A weighted average of precision and recall.
Feature engineering	The use of features of data by a machine-learning algorithm to achieve specific tasks such as mapping.
Granularity	The scale or level of detail present in a set of data.
Information retrieval	The science of searching for information of unstructured nature in a document, searching for the document itself, and

	searching for the metadata that describes data. This study uses the term to refer generally to finding information of interest.
Knowledge organization system	Knowledge Organization Systems cover a wide range of items (subject headings, thesauri, classification schemes, and ontologies), distinguished by their specific structure and function and used in diverse contexts to support the organization of knowledge and information to facilitate information management and retrieval.
Knowledge representation	The study of how an intelligent agent's beliefs, intentions, and value judgments can be expressed in a transparent, symbolic notation suitable for automatic reasoning.
Lexical text matching	Text matching that is based on the level of words concerning their lexical meaning and part-of-speech.
Linked data	A method for publishing structured data using vocabularies that can be connected and interpreted by machines.
Mapping	The directed alignment of entities of one ontology to at most one entity of another ontology.
Meaningful Use	A term used to define minimum U.S. Government standards for electronic health records (EHR), outlining how clinical patient data should be exchanged between healthcare providers, providers and insurers, and providers and patients.

Morphological text matching	Text matching that exploits word structures and word formation, focusing on analyzing the individual components of words.
Natural language processing	An area of artificial intelligence research that explores ways to automatically understand and manipulate natural human language such as that contained in speech and text to perform useful tasks.
Ontologies	Type of KOS defined as an explicit specification of a conceptualization, a representational vocabulary for a shared domain of discourse (definitions of classes, relations, functions, and other objects), i.e., a model for describing the world that consists of types, properties, and relationships.
Precision	Also referred to as positive predictive value, is a measure of the fraction of relevant instances among the total retrieved instances. In mapping, it is a measure of the fraction of system assignments made that are correct.
Recall	Also referred to as sensitivity, is a measure of the fraction of relevant instances retrieves over the total amount of relevant instances. In mapping a measure of the fraction of total word instances correctly assigned.
Reference terminology	Sets of concepts and relationships that provide a common reference point for comparing and aggregating data about the healthcare process.

Relations	Ways in which concepts or entities can be related to one another.
Semantic analysis	A method for minimizing syntactic structures and providing meaning, finding synonyms, word sense disambiguation, translating from one natural language to another, extraction of entities and relations, and populating knowledge base.
Semantic enrichment	Enhancement of content with information about its meaning.
Semantic equivalence	A declaration that two data elements from different vocabularies contain data that has a similar meaning.
Semantic interoperability	The ability to use digital health information across diverse settings and clinical software as increasing amounts of health data from diverse locations makes for unique challenges in connecting and analyzing these data as a unified set.
Specificity	Refers to the amount of domain-specific information present in a term
Structured data	Refers to any organized data that resides in a fixed field within a record or file in a certain format.
Supervised learning	Machine-learning algorithms that learn by example input and output are used to train the model to make its inferences.
Terminologies	

	Terminologies are products of science that aim to make an inventory of given domain concepts and terms that designate them.
Unstructured data	Refers to data that does not conform to the data model nor has any structure.
Unsupervised learning	A machine learning algorithm used to make inferences from datasets consisting of input data without labeled responses.

Chapter 1. Introduction

In March 2020, the world became broadly aware of a threat to humankind that had been quietly brewing for several months. The coronavirus disease 2019 (COVID-19) pandemic has revealed challenges and opportunities for data analytics, semantic interoperability, and decision making. The sharing of COVID-19 data has become crucial for leveraging research, testing drug effectiveness and therapeutic strategies, and developing policies for control, intervention, and potential eradication of this disease. Sharing and assessing accurate and detailed clinical data is critical and yet one of the more difficult challenges in dealing with the pandemic.

In the past decade, especially in the United States, healthcare policymakers have brought attention to the need for electronic health records, information exchange, and interoperability of health systems, citing the aims of improving patient safety, reducing medical error, improving efficiency, and reduction of cost. Furthermore, as other medical informatics applications are developed, and the potential for secondary use of data hidden in medical documentation and clinical trials is realized, the need for integrated clinical coding schemes increases exponentially. Health information systems must represent the findings, management, and outcomes of the patients. Ideally, they should do this while preserving clinical detail, identifying characteristics for improving risk, aggregating outcome analyses, and enabling decision support (Chute et al., 1999) through the use of clinical coding schemes which specify and define concepts in a domain.

Clinical coding schemes help achieve meaningful and accurate exchange and use of information, enriching knowledge and facilitating improved information analysis (Arvanitis, 2014; Zeng et al., 2020). They further enable the capture of clinical findings, natural language processing, indexing medical records and literature, representing medical knowledge, and more (Cimino, 1998). Clinical coding schemes, used as a term in this document to broadly represent structured vocabularies, classification schemes, and ontologies in the biomedical domain,

function as the backbone of core processes that often occur in the medical field today and must meet high expectations from the health care community. They are critical for defining and structuring concepts and terms in healthcare to ensure consistent use by stakeholders within the industry. Clinical coding schemes equip knowledge organization systems with various abilities to support health care. For example, they support data sharing, link clinical evidence with administrative decisions, support evidence-based practice, enable population-based interventions, use electronic health records and decision support systems, and advance medical research.

Deficiencies in the healthcare systems such as inadequate patient information at the point of care, flawed and misleading data that result in disorganization, and errors in clinical care and administration are often the result of poor implementation of standards for format, content, language, and completeness (Rose et al., 2001).The implementation of these standards within the healthcare system itself can often be problematic. Healthcare concepts often have multiple identifiers and descriptors within and across systems, resulting in clinical misinterpretation, inadequate or incorrect knowledge management, and misdiagnoses of patient's problems. Vocabulary developers have responded by adding even more terms and offering new, improved versions of their coding schemes, yet this is not enough. Estimates of the number of terms needed to describe health-related concepts place the number at about 45 million, covering concepts related to medicine, biomedical molecules, genes, organisms, patients, conditions, populations, healthcare actions, technical methods, and social concepts (ISO, 2018). Agreeing on standard terms and establishing reliable terminology can improve the semantic interoperability of information in disparate systems.

These issues are further complicated by the important information hidden in unstructured form within medical records, clinical or laboratory reports, patient notes, and free-text responses sections of case report forms and clinical trials. Clinical trials are used to gather safety and efficacy data on new drugs in development or the use of existing drugs in new contexts. Some

information is structured and already searchable with keywords, but questions remain about the accuracy and completeness of the coding. Further, much of the information contained within these documents lies in portions of unstructured text, which are not coded with clinical coding schemes at all. While natural language allows for rich and detailed documentation, it suffers from ambiguity due to its dependence on contexts, jargon, acronyms, and lack of strict definitions. Conversely, structured data constrains expressiveness and flexibility and increases the difficulty of interpreting or recreating meaning due to contextual information loss. Thus, there is a need for a "common, uniform, and comprehensive approach" to clinical knowledge representation (de Quiros et al., 2018).

Providing the best care to patients depends on assessing qualitative, unstructured data, which is often subjective and specific to the patient but aids greatly in making correct diagnoses or achieving a successful drug approval process (Smithwick, 2015). Taking steps to provide semantic annotation of unstructured data enhances discovery, interpretation, and reuse. Annotated data allows easy detection of equivalent concepts, disambiguation of terms, and the allowance of hierarchical searches. It further provides a machine-readable HTTP URI that resolves and dereferences to a helpful specification of other relationships for that annotated resource (Jones et al., 2019). When data is structured, meanings are consistent, and it can be searched with algorithms and ontologies which can infer context. Semantic technologies applied to unstructured data allow machines to process data more quickly, providing benefits to both researchers and patients. Therefore, the high-value information stored in unstructured form needs to be extracted and synthesized. Smithwick (2015) indicates this is done by creating "structured symptoms, i.e., gathering the information in the unstructured portions and discretely capturing them in a way the data can be analyzed," which can only be achieved through the application of clinical coding schemes.

1.1 Clinical Coding Schemes

Clinical vocabularies, terminologies, or coding systems, which in this work are referred to collectively as clinical coding schemes, are structured lists of terms and their associated definitions intended to describe the healthcare domain categorically. Clinical coding schemes can be defined as standard terms or synonyms that record patient information to support clinical care, decision support, outcomes research, and quality improvement (Chute, 2000). They are part of a class of objects known as Knowledge Organization Systems (KOS). Zeng (2008) explains that KOS can be organized into four main groups ranging from simple to complicated. These are Term Lists which include pick lists, dictionaries, and synonym rings. Metadata-like models, which include authority files and directories. Classification and categorization, which includes subject headings, taxonomies, and classification schemes. And finally, relationship models, including thesauri, semantic networks, and ontologies, are shown in Figure 1.

The NIH has mandated the adoption and use of clinical coding schemes such as Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC) RxNorm. There seems to be a lack of agreement on exactly what to call these coding schemes in the medical literature. However, what is common is that these schemes function to eliminate ambiguity, control synonyms, establish hierarchical and associative relationships, present properties, and represent the underlying semantic structure of a domain (Zeng, 2008). These schemes represent diseases, diagnoses, treatments, findings, operations, observations, medications, administrative concepts, and more in the clinical domain (OpenClinical, 2005).

Figure 1

Overview of the Structures and Functions of KOS

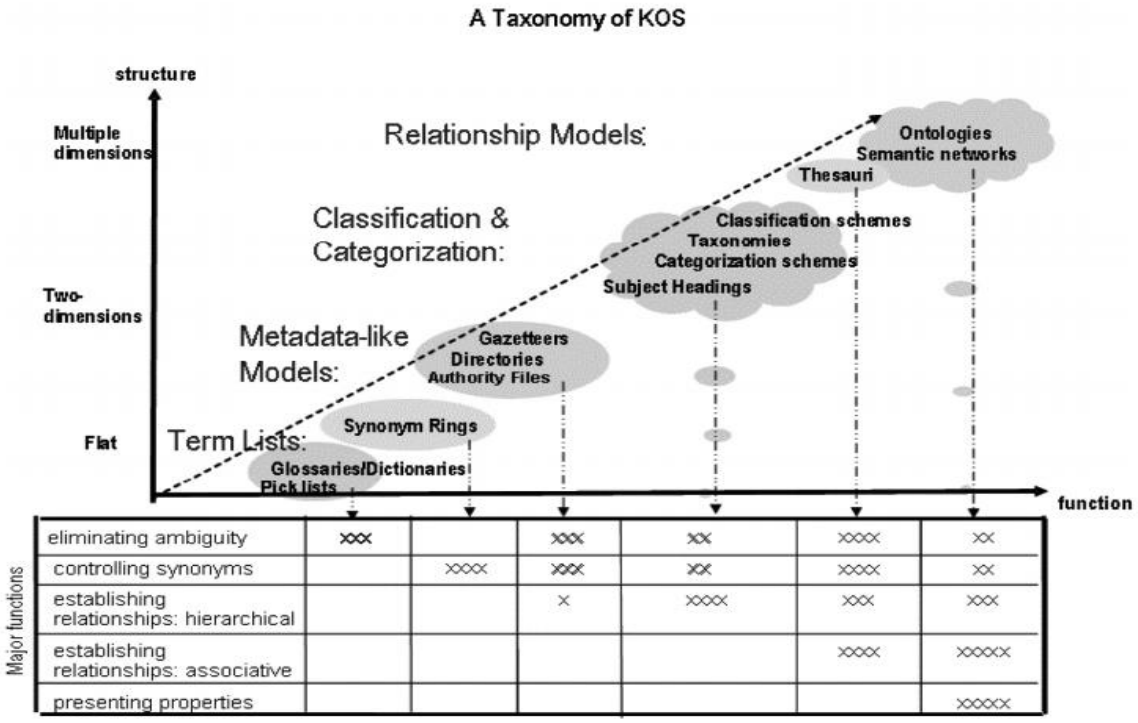


Figure 1. *An overview of the structures and functions of KOS*

Note. Adapted from (Zeng, 2008)

Various expressions have been found, including controlled health terminologies, clinical terminologies, clinical classification systems, healthcare terminologies, standard terminologies, controlled medical terminologies, biomedical terminologies, and the like. Still, others are being developed and used in specific contexts. A brief overview of some of these follows.

1.1.1 Example Clinical Coding Schemes.

1.1.1.1 ICD-10. The International Classification of Diseases (ICD) is a standard classification system developed by the World Health Organization (WHO) for hospital diagnosis, procedure billing and encoding for clinical use, health management, and epidemiology. The ICD defines the universe of diseases, disorders, injuries, and other related health conditions in a comprehensive, hierarchical fashion. It is used to compile health statistics, monitor spending,

inform policy, outcome prediction, and analyze the general health situation of population groups. In addition, it is used to surveil the incidence and prevalence of diseases and other health problems.

1.1.1.2 LOINC. Logical Observation Identifiers Names and Codes (LOINC) was created in 1994 and is maintained and distributed by the Regenstrief Institute with support from the National Library of Medicine (NLM). The LOINC database provides a universal code system for laboratory reporting and other clinical observations. Many laboratories and clinical services use HL7 to send results electronically from their reporting systems to their care systems. LOINC functions as a common language, i.e., a set of identifiers, names, and codes that function to identify health measurements, observations, and documents. It applies a universal identifier to medical terminology related to electronic health records. It enables the exchange and aggregation of clinical results for care delivery, outcomes management, and research through these codes, which allow for the structured names which remove ambiguity in identifying concepts that can be measured or observed.

1.1.1.3 RxNorm. RxNorm is a system that provides normalized names and unique identifiers for generic and branded drugs and a tool that enables semantic interoperability between drug terminologies and pharmacy knowledge base systems. It is made available by the National Library of Medicine (NLM) and is used by hospitals, pharmacies, and other organizations to process and record drug information. RxNorm derives its drug names from multiple data sources (DrugBank, Medical Subject Headings, Multum MediSource Lexicon) commonly used in pharmacy management and drug interaction software (NLM, 2020). It preserves the meanings, names, relationships, and attributes from the sources. In the RxNorm drug model, normalized names are organized in a pattern of ingredient, strength, and dose form. It also includes two additional elements, quantity factor, and qualitative distinction. Information such as indications, drug classes, drug-drug interactions, and drug pricing is not included in RxNorm. However, it does integrate codes from the National Drug Code, which

serve as product identifiers for drugs in billing transactions. RxNorm focuses mostly on drug products marketed in the USA despite its integration of international sources. RxNorm has been used in various applications such as electronic prescribing, information exchange, formulary development, reference value sets, and analytics (Bodenreider et al., 2018).

1.1.1.4 SNOMED-CT. The Systemized Nomenclature of Medicine – Clinical Terms (SNOMED CT) is a controlled clinical reference terminology with comprehensive coverage of diseases, clinical findings, etiologies, procedures, living organisms, and outcomes used by clinicians, including physicians, dentists, nurses, and allied health professionals in recording and documenting patient data. SNOMED CT is one of the standards designated by the U.S. government for the electronic exchange of clinical health information and is one of the required standards for interoperability specified by the U.S. Healthcare Information Technology Standards Panel (NLM, 2019). In the U.S., clinicians must encode problem lists, procedures, and other concepts using it to meet Meaningful Use Stage 2 requirements. Meaningful Use requirements cover maintaining up to date problem lists of current and active diagnoses, recording patient family health history as structure data, identifying and reporting cancer cases to state cancer registries, recording and tracking changes in patient vital signs, recording patient smoking status, and providing summary records for care transitions.

1.1.1.5 COVOC. The abbreviation COVOC represents COVID-19 Vocabulary, an ontology containing terms related to the research of the COVID-19 pandemic such as host organism, pathogenicity, gene and gene products, barrier gestures, treatments, et cetera(EMBL-EBI Ontology Lookup Service, 2021).

1.1.1.5 CIDO. The Coronavirus Infectious Disease Ontology is an open-source biomedical ontology for coronavirus infectious diseases. It was developed to provide standardized human and computer interpretable annotation and representation of various coronavirus infectious diseases, including their etiology, transmission, pathogenesis, diagnosis, prevention, and treatment (National Center for Biomedical Ontology, 2021).

1.1.1.6 COVID-19 Ontology. This ontology contains concepts covering the role of molecular and cellular entities in virus-host interactions, in the virus life cycle, as well as a wide spectrum of medical and epidemiological concepts (National Center for Biomedical Ontology, 2021).

1.1.2 Summary

Many coding schemes for the healthcare domain have been developed, and some have been recommended for adoption. These schemes are almost in a state of competition with each other, not least because their coverage and content are so varied. In many cases, they overlap each other, although they are designed to meet a variety of well-defined goals. Because these schemes are either not detailed enough, focus on a particular narrow domain of healthcare, are proprietary or custom-built, or just difficult to use, achieving semantic interoperability remains a persistent challenge. It would be ideal if a single, integrated, and comprehensive scheme could meet the needs of all.

1.2 Rationale

As indicated previously, maximizing the reuse of data has become increasingly important in healthcare. However, the data description has often been lacking in various ways, impeding advancement in enabling semantic interoperability, health information exchange, analytics, and research. Further, data stored in siloed systems cannot interact with other systems at the semantic level. The number of terminologies and the lack of consistent or standard usage across applications impede data sharing and aggregation, making it difficult for systems to communicate and increasing the challenges faced by clinical professionals and researchers alike. Clinical coding schemes are a crucial element of the infrastructure needed for enabling the proper functioning of healthcare systems, particularly for facilitating data-driven research discoveries (Schriml et al., 2020). In an age of ever-emerging new diseases and

healthcare challenges, for example, COVID-19, the necessity for expanding the reusability of data becomes increasingly apparent.

Researchers may struggle to find answers to the fundamental questions they are interested in due to variations in the amount of concept information represented in medical terminologies or the lack of applied standards describing the data. They may encounter problems caused by a lack of mapped data, semantic harmonization, and terminology integration. Due to the vast amounts of data generated, many documents and applications require multiple linked data sources to gain the most value from them. Translating healthcare data between various types of core reference terminologies used to describe patient data, reporting, administrative or epidemiological classification purposes such as billing or mortality reporting is often necessary. Applications that involve multiple coding schemes must establish semantic mappings among them to ensure interoperability.

The FAIR principles (Wilkinson et al., 2016) outline the need for infrastructure that supports data reuse through processes that enhance a machine's ability to automatically discover, use, and reuse data by making them findable, accessible, interoperable, and reusable. Further, this infrastructure should be made functional, impactful, and transformable (FIT) to truly function as the critical components needed for acting as the framework needed to support data-driven and AI-dominated processes (Zeng & Clunis, 2020). These infrastructures rely on concepts that often have multiple identifiers and descriptors. Therefore, a standard and reliable coding scheme must be achieved to improve semantic interoperability in disparate systems.

Critically, because no single code set has managed to meet all medical institutions' needs, various efforts have been made for integration. Mapping from one coding scheme to another is often difficult to accomplish for a variety of reasons. These include the many-to-many mappings common between terminologies, the similarity between concepts in one scheme making it difficult to map to another, or incomplete mapping rules or issues with granularity making selecting appropriate codes difficult.

Accurate mappings between clinical coding schemes function to improve efficiency and promote better sharing, combining, and linking data sets from different sources and ensuring that the meaning of information coming from disparate systems is the same. Further, it allows comparisons between research studies which would otherwise be impossible because of confusion caused by lack of alignment (Gliklich et al., 2014). Mappings between coding schemes will be critical for helping organizations that still have legacy data move it into the future and support browsing and searching of unstructured data such as clinical trials through semantic annotation.

Aligning terminologies through mapping supports information retrieval and integrates data from different resources into a single context to enhance understanding of complicated biomedical systems. Mapping challenges could lead to claim rejection due to insufficient documentation and lack of evidence or affect clinical decision-making because of the inconsistencies between health problems and treatment plans. Furthermore, the description of concepts for new diseases and alignment of those terms with preexisting terminologies is a current and pressing issue. Failing to enhance clinical coding schemes through mapping or linking between terminologies is a serious hindrance to the research needed in medical crises, such as with the current pandemic.

Zeng (2019) outlines various challenges to mapping involving the structure, domain, language, or granularity of coding schemes. In addition, many of the current methods for mapping/alignment are heavily manual, time-consuming, and error-prone, and challenging at web-scale resulting in serious detrimental consequences for clinical misinterpretation, mediocre and incorrect management of knowledge, or misdiagnoses of patient's problems. Zeng (2019) defines mapping as the process of establishing relationships between the contents of one vocabulary and those of another. Therefore, for interoperability to be a reality, data integration through mapping will be critical in delivering quality services as data is being ingested, captured,

and collected from multiple sources. The integration and interoperability of these resources are key to enabling applications that will answer questions that currently elude us.

1.2.1 Semantic Interoperability

The problem of interoperability is one of the main clinical and research challenges in healthcare today and particularly secondary use of clinical data. Semantic interoperability describes computer systems' ability to exchange data with unambiguous, shared meaning required to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems (Geraci et al., 1991)(Geraci et al., 1991). Given the number of different standards or data formats used by different databases participating in biomedical and clinical sciences research, translating into an intermediate, common format or standard for use within the network offers an opportunity to reduce translations, thus providing greater efficiency and simplifying scalability.

Within the context of clinical coding schemes, certain interoperability issues are likely. These are differences in encodings and representations (syntactic), variances in data models, data structures, and schemas (structural), and inconsistencies in terminologies and meanings (semantic) (Zeng, 2019). Arvanitis (2014) further expands this idea to explain that the syntactic level is concerned with the standardization of data formats and communication protocols and provides the basic links and integration between systems and components, enabling information exchange. In contrast, the semantic level aims to develop user understandable, computable, and extensible knowledge representation schemes to capture concepts and information usable by machines and humans.

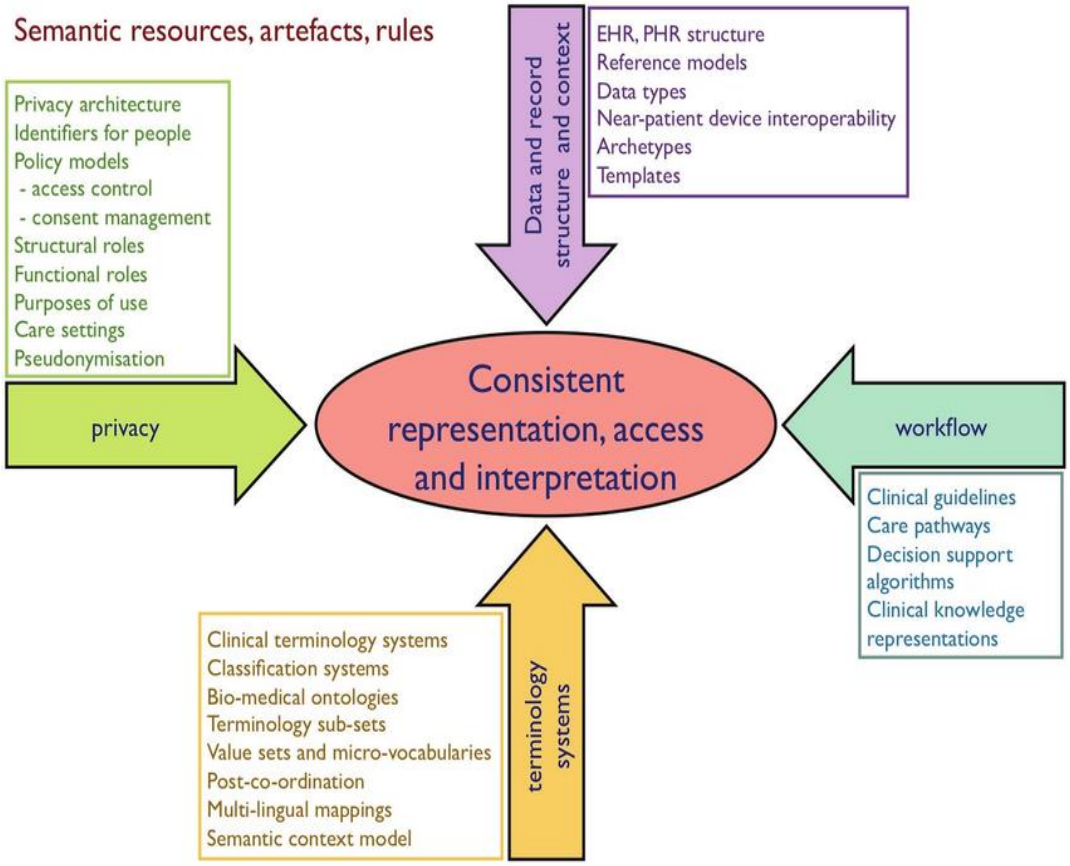
These knowledge representation schemes help achieve meaningful and accurate utilization of the information exchanged at the syntactic level of interoperability and further act as a method for information enrichment and facilitate better information analysis processes (Arvanitis, 2014). These schemes are critical for creating insight and bridging the contextual differences across systems (Zeng, 2019). For example, consider the semantic interoperability

resources shown in Figure 2, which shows the various resources that need to be integrated to get the most value from an electronic health record, which requires interoperability among clinical coding schemes. In a healthcare system, semantic interoperability enables digital health information across diverse settings and clinical software. Increasing amounts of health data from diverse locations make for unique challenges in connecting and analyzing them as a unified set.

In the healthcare systems, various standards are employed for different services. For example, it is difficult to integrate and exchange medication information since systems often use different terminologies. A pharmacy might use a formulary service terminology while the Computerized Physician Order Entry (CPOE) system uses another terminology. Such terms can be even further modified at different points in the system to achieve consistency with naming conventions used by the Federal Drug Administration (FDA) or the National Drug Code (NDC). Clinical coding schemes can help to eliminate semantic conflicts and enhance information exchange and communication. Thus, there is added value in designating healthcare concepts to meaningful descriptive terms, associated coding systems, and supportive vocabulary services to achieve semantic interoperability within the healthcare context (Arvanitis, 2014). An integrated coding scheme must be leveraged for systems with diverse data sources and coding schemes. Although there are growing collaborative efforts between clinical coding scheme developers to improve compatibility and extensibility in clinical coding schemes, researchers still contend that mappings between the coding schemes are required. Mappings are critical since the formalisms and tools used for representation in each, together with the release cycles and versioning mechanisms, decrease the likelihood of seamless integration that is the objective of these collaborations (Bodenreider et al., 2018).

Figure 2

Semantic Interoperability Resources



Note. Adapted from (Kalra et al., 2011). ARGOS Policy Brief on Semantic Interoperability.

Most attempts at solutions to these issues use problem-specific algorithms that are labor-intensive, difficult to maintain, or unscalable outside of the domain where they were first deployed. Finding an approach that automatically identifies relevant biomedical concepts across coding schemes while requiring less labor is easily maintained and replicated is a project worth exploring. Semantic mappings can likely be identified using automated methods or through an approach that uses clinical terminologies' semantic or structural properties when mapping them

to each other. These may perform better than manual approaches in classification, accuracy, computational time, and scalability.

1.3 Objective

The previous discussion on the challenges to mapping, semantic interoperability, and access to information stored in unstructured documents highlights the need for tools, processes, and methods to address the issues outlined. As the world deals with the challenges of a new disease, we must take every opportunity that leads to new knowledge discovery. Given the decentralized nature of the clinical coding schemes and systems involved and the expected continuous explosion in their numbers, tools that focus on these issues and support the pandemic response are needed.

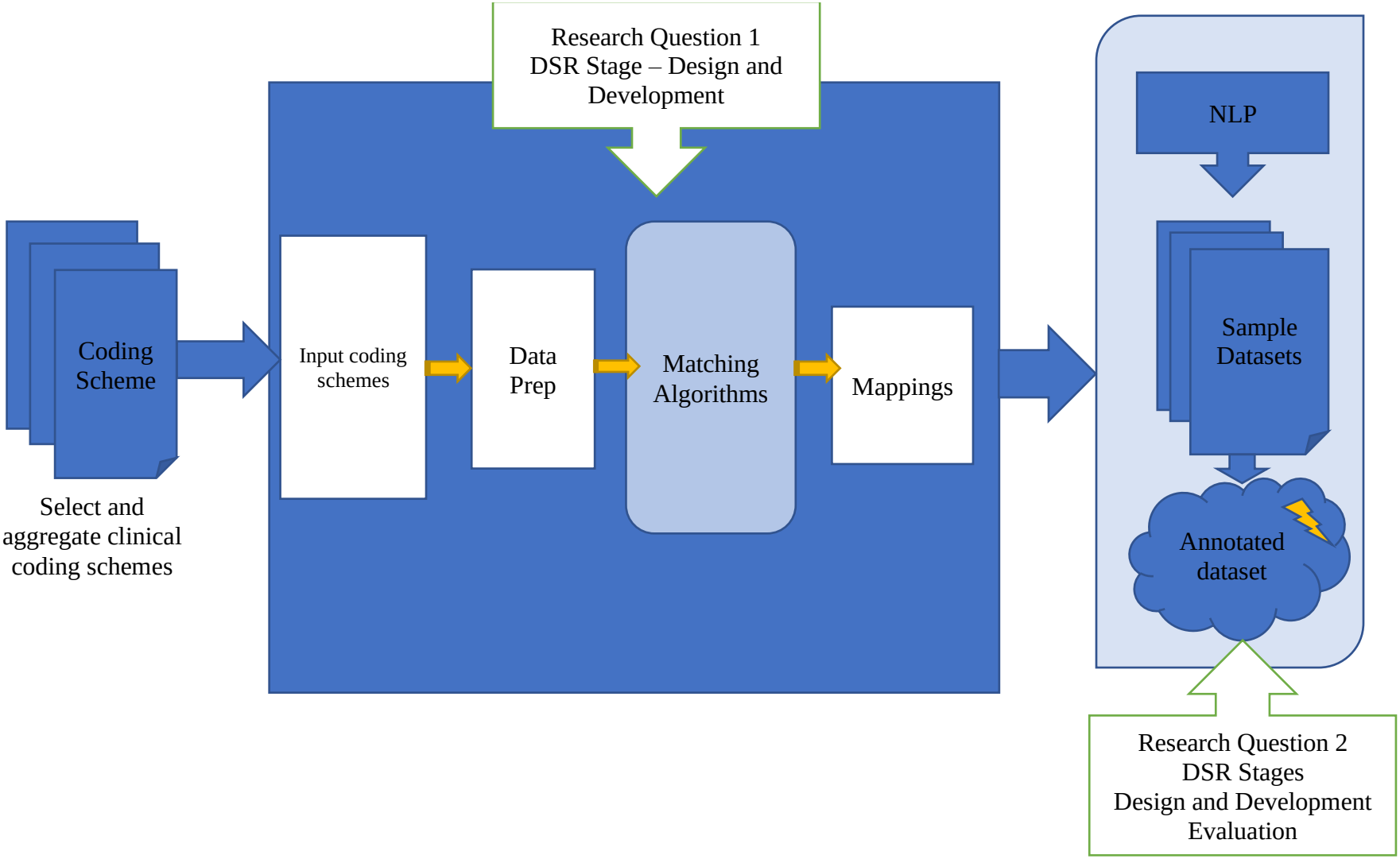
In the past year, as the world has dealt with the pandemic, several terminology resources are in development to respond to the unique terminology needs of the current COVID-19 pandemic. Examples include the Coronavirus Vocabulary COVOC, Coronavirus Infectious Disease Ontology (CIDO), the COVID-19 Surveillance Ontology, and the WHO COVID-19 Rapid Response Version CRF Semantic Data Model (COVIDCRFRAPID). Most of these coding schemes have been made available as ontologies. They can be accessed through registries such as BioPortal or the European Bioinformatics Institute's Ontology Lookup Service. Some of these terminologies have limited mappings to other vocabularies, such as LOINC. They have not yet been mapped to each other, and some have no mappings to other coding schemes at all. Additionally, since these are newly developed, most clinical trial documentation has not had structured or unstructured data fields coded with these.

My objective then is to develop a tool in the form of a reusable workflow to help healthcare stakeholders take advantage of the clinical coding schemes available for COVID-19 with mappings to other current medical standards such as LOINC and SNOMED. Creating mappings for these new ontologies could help support the work being done by researchers using them. In addition, the mapped terms will be used to support semantic annotation of clinical

documents that deal with COVID-19. The expectation is that a unified clinical knowledge representation approach would positively impact health determinants, long-term prognosis after diagnosis and intervention, and research advances. A conceptual diagram of the anticipated workflow is presented in Figure 3.

Figure 3

Conceptual Model of Semantic Analysis Framework



1.3.1 Research Questions

Based on the problems and objectives outlined, the following research questions will be addressed:

Research Question 1. How can an Extract Transform Load (ETL) workflow tool support the task of clinical coding scheme mapping?

Research Question 2. How does the mapping output of the novel workflow support and affect annotation of clinical trials in COVID-19 research?

Research Question 3. What aspects of the sociotechnical model can be leveraged or updated to explain and assess mapping to achieve semantic interoperability in clinical coding schemes?

1.3.2 Discussion of Research Questions

Extract Transform and Load tools offer critical functionality to people wishing to wrangle data in multiple formats where information exists but making sense of it is difficult. These tools offer functionality similar to business intelligence tools and can perform tasks from data blending to predictive analytics and produces useful output, visualizations, and even dashboards. Usually, vocabulary integration, alignment, mapping, and annotation tasks are complicated by their heavily manual, resource, and time-intensive nature. Often the process requires technical knowledge involving multiple individual experts and software tools. This study investigates the use of these tools for mapping, evaluating their functioning, and whether they offer improvements over traditional methods such as using an ontology alignment tool or manual mapping of codes.

A whitepaper published by Antitdote (2021) indicated that searches for clinical trials increased by 22% in March 2020 compared to March 2019. Their in-house clinical research trials have seen engagement rates increase by 27%, registration rates by 43%, and their need to advertise for participants decrease by 53%. These statistics support the idea that data

volumes are increasing. As outlined previously, clinical trials contain unstructured elements that contain information useful for medical research. Taking the best advantage of this data requires annotating text with terminologies and ontologies (Tchechmedjiev et al., 2018). In addition, with the increase in complexity and volume of COVID-19 clinical documentation, it would be useful to extract all potential points of data contained in those texts. Semantic annotation of this data can facilitate mapping the data elements to diverse sources, supporting data integration, decision support, and surveillance.

Finally, clinical coding schemes and the mapping process itself exist within the context of sociotechnical systems. Additionally, many of these tools have complicated requirements, are built for specific use cases, are proprietary, and have high costs of both time and finances to implement or are beyond the scope of expertise of the stakeholders. The challenges of mapping and maintaining those mappings are a significant task beyond a human/s ability to handle alone. Thus, as in any socio-technical system, semantic interoperability needs the coordination of people, processes, and tools.

The sociotechnical model has typically been used to assess the design, development, implementation, use, and evaluation of health information technology within complex healthcare systems. It often addresses individuals' characteristics, work tasks, physical environment, human-system interfaces, and organization. An exploration of mapping through the lenses of the human, social, technological, and organizational elements of the entire healthcare process and consider impacts, revisions, and updates based on the knowledge gained through the design of the novel workflow is undertaken.

1.4 Relevance and Significance

The literature review will show that many researchers have explored the issue of semantic interoperability (Arvanitis, 2014; Binding & Tudhope, 2016; Dias et al., 2014; Kalra et

al., 2011; Zeng, 2019). Yet although the problem has been researched from multiple perspectives, the consensus is that health information exchange is still challenging.

1.4.1 Health Information Exchange (HIE)

The Office of the National Coordinator for health information technology (2019) defines health information exchange as a means for allowing clinicians to access and securely share medical information electronically appropriately. Health information exchange enables the interoperability of automated health data to support improvements in healthcare quality and efficiency (Kuperman, 2011); improve population health and improve the emergency response (Shapiro et al., 2011); lower costs across health systems and improved patient safety (Menachemi et al., 2018). Semantic interoperability makes health information exchange possible as it allows for the synthesis of codes from multiple coding schemes.

Several clinical coding schemes have been developed, mandated for implementation, or created for specific contexts. In much the same way, researchers have risen to the challenge of creating coding schemes for sharing COVID-19 data. This proliferation of schemes has contributed to the problems identified in this study. There is no comprehensive standard that can meet the demands of data exchange for clinical professionals and researchers. Therefore, this research is relevant to the goal of providing an interoperable solution for data exchange. The literature review will also highlight the lack of simple methods or a single method for performing mapping tasks. The research solution – a reusable novel workflow tool for mapping clinical coding schemes and annotation of clinical trials – will add to the body of knowledge an artifact that can support interoperability.

1.5 Ethical Approval

Because this research analyses existing publicly available data relating to clinical trials and their characteristics rather than human participants, ethical approval is not required for this research.

Chapter 2. Literature Review

To gain the most value from data, facilitating data sharing, information retrieval, interoperability, and appropriate annotation and classification of clinical trials, different terminology sets, and subject/coding schemes must be linked to one another through the mapping process. Mapping is an effective and widely used approach for semantic interoperability based on creating links between different coding schemes. Mapping also removes barriers resulting from multilingual schemes (Zeng, 2019). However, there is an extensive time and resource commitment necessary to accomplish mapping, especially for schemas with varying degrees of granularity, making automated mappings more complex (McCulloch et al., 2005). Some challenges come from the theoretical, conceptual, cultural, and practical differences, mapping terms of different hierarchical status and specificity levels, or the need to update mappings when new versions of coding schemes are released.

2.1 A Theoretical View of Clinical Coding Schemes and Mapping

2.1.1 Social Construction

Terminology development is a socially important activity. It is the discipline concerned with the study and compilation of specialized new terms, and it has social and political implications. As science became a worldwide phenomenon, the need grew for scientists to have rules for formulating terms in their fields. Edwards (2004) suggests that standards are socially constructed tools that embody negotiations between the technical, social, and political. Standards enable the construction of technological systems by making it possible to disseminate knowledge (Edwards, 2004).

Cultural changes have occurred regarding the value attached to information as technology becomes more casually prevalent in society. As products, services and knowledge became more widely exchanged, the need to standardize elements of scientific, technical,

cultural, commercial, and biomedical domains increased (Edwards, 2004). Technological change has spurred scientists worldwide to create hierarchical rules for good usage of terms describing artifacts and domains, giving rise to standards. Further highlighting this need, government mandates make standard creation a necessary endeavor (AHIMA and AMIA Terminology and Classification Policy Task Force, 2009; Institute of Medicine (US) Committee on Data Standards for Patient Safety et al., 2004). Edwards (2004) notes that the dominant economic powers influence scientific and technological creation enabling one-way transfers of knowledge and necessitating borrowing of information facilitated through standard terminologies. Because standards are designed to work in one way regardless of the circumstance, they can be built into systems.

In the healthcare context, clinicians have realized the need for standard terminologies to aid them in the exchange of information, to be able to describe observations, diseases, diagnoses, and other clinical terms in standard ways. However, with the commercialization of health care and the involvement of insurance companies and big government came policies and mandates that influence what standards can be applied. Those decisions determine the 'status' of a hospital as far as meeting meaningful use requirements goes. They further determine payment and reimbursements, which in turn has impacts on decision-making. The standards that are mandated and built-in to EHRs e.g., SNOMED-CT have had significant impacts on clinician workflows and, in turn, their experience of the workspace and also has an impact on patient experience and safety.

In addition to these mandates, clinicians must contend with a growing amount of diverse information objects, changes in technology, and the need to have immediate, reliable, stable, and comprehensive access to information. To meet these demands, semantic enrichment of information objects supported by clinical coding schemes must occur (Alemu et al., 2012). A socially constructed approach to mapping might allow users access to the content and the ability to participate in the process of creating it. Participation might entail selecting mapping

terms, recommendations for what to include in a tool or process, or even machine-generated terms from social network content. The issue with this approach is that users are not always aware of the constraints imposed by established standards, and so there may be a disconnect between what they think they need or how they describe concepts and what established systems require. Mapping could provide common ground by allowing links between socially constructed terms and standard terminologies in a way that would benefit all.

2.1.2 Sociotechnical Systems Theory

Design, development, implementation, and evaluation of Health Information Technologies (HIT) continues to be one of the major challenges within the health care system. Various conceptual models of user interaction with technology, use, acceptance, and evaluation have been created to understand this issue. These include Roger's diffusion of innovation theory (Rogers & Marshall, 2003), Venkatesh's unified theory of acceptance and use of technology (Venkatesh et al., 2003), Hutchin's theory of distributed cognition (Hutchins, 1995), Reason's Swiss Cheese Model (Reason, 2000) and Norman's 7-step human-computer interaction model (Norman, 1988). However, these models do not address the entire range of issues that must be considered when designing, developing, implementing, using, and evaluating HIT.

Sittig and Singh (2015), in a review of the models above and their application to healthcare, suggest these models do not do enough to consider the relationships that exist between hardware, software, information content, and the human-computer interface in the healthcare context. Health care is happening in various physical and organizational settings and environments that are either loosely or tightly connected (Carayon et al., 2011). These connections are often enabled and supported through the clinical coding schemes built into the system that describe clinical care contexts.

Sociotechnical models deployed in health care contexts attempt to address these issues by treating HIT-enabled healthcare systems as complex adaptive systems. Early

implementations of the sociotechnical systems model in healthcare focused on a small subset of facets. For example, Henricksen and Kaye (2003) focus on provider characteristics, the attributes, and difficulty of tasks, the environment in which it happens, the human-system interfaces involved, and its characteristics. Carayon (2006) focused on characteristics of providers such as tools, resources, organization settings, interpersonal and technical aspects of health care activities, and changes in patients' health status and behavior.

One healthcare model that focuses on the individual components of technology deserves a closer look. This model allows implementation and usage problems to be more easily identified and specific solutions created (Sittig & Singh, 2015). The authors particularly highlight the role of controlled clinical vocabularies that act as a "cognitive interface between the inexact, subjective, highly variable world of biomedicine and the highly structured, tightly controlled, digital world of computers." Noting their potential impacts when not distinguished and addressed specifically. They outline a new sociotechnical model for HEALTH IT that involves eight dimensions. These steps are neither independent nor sequential but instead interact with each other in various ways that should be assessed. The eight dimensions are "1) hardware and software infrastructure, 2) clinical content, 3) human-computer interface, 4) people, 5) workflow and communication, 6) internal organization policies, procedures, and culture, 7) external rules, regulations, and pressures, and 8) system measurement and monitoring” (Sittig & Singh, 2015).

Thinking of these dimensions relative to applications and workflows for scheme matching, one can see that a sociotechnical view of the mapping process would necessarily consider the applications used to enable mapping. It would also consider the clinical coding schemes being aligned, the interface through which users interact with the tool, and the people who design, test, and use it. Furthermore, the mapping tool's impact on the workflow of clinicians, coders, and researchers who might use it and the policies that have made it necessary to perform scheme mapping would also be addressed. In addition, it would consider

the impacts policies, procedures, and culture have on the kinds of decisions made during the design and development process of a mapping solution.

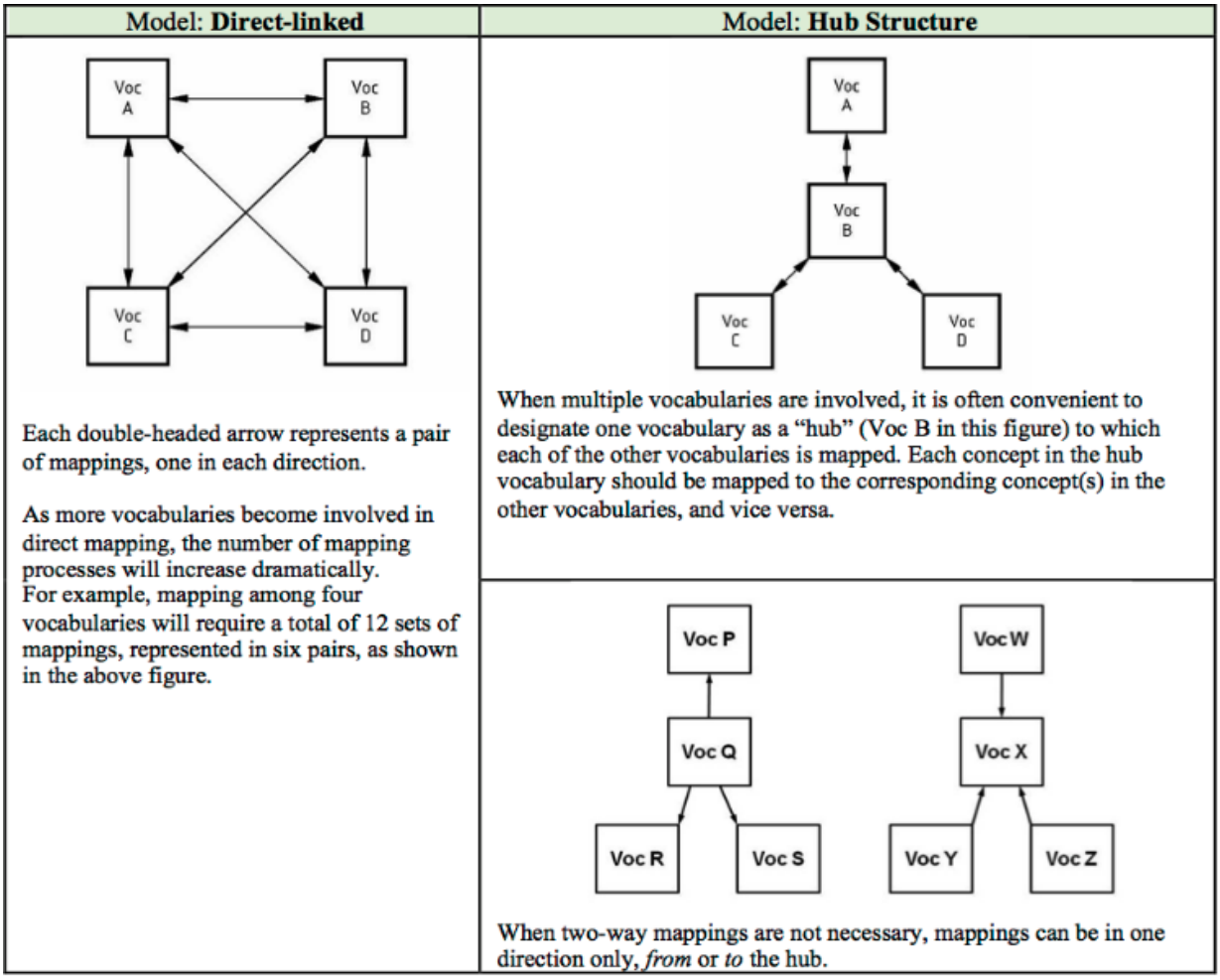
2.2 Mapping Approaches

Theoretical perspectives can guide our understanding of the factors influencing the creation of clinical coding schemes and the reasons why it is necessary to create links between them. It also highlights the challenges to enabling interoperability between the complex and competing systems involved in health care processes. Chute (2000) describes how some developers and authors discuss terminologies in terms of competing with one another and not having a specific role to play. Rather than focus on differences, attention should be given to enabling interoperability through linking these schemes.

Mapping is not a new problem, and some guidelines exist for how to address the issue. Zeng (2019) mentions two mapping structures recommended by the ISO 25964-2:2013 for vocabularies that do not share the same structure, scope, language, or typology outlined in Figure 4 and highlights the fact that these mappings require significant work to build and maintain. Other interesting mapping methods outlined in Zeng's work include selective mapping where mappings are applied only for the concepts that have been used or are likely to be used within the application in question; cooccurrence mapping, which leverages social network platform information; and blended mapping where multiple models are used in the same case depending on the situational contexts.

Figure 4

Mapping Models for Dissimilar Vocabularies

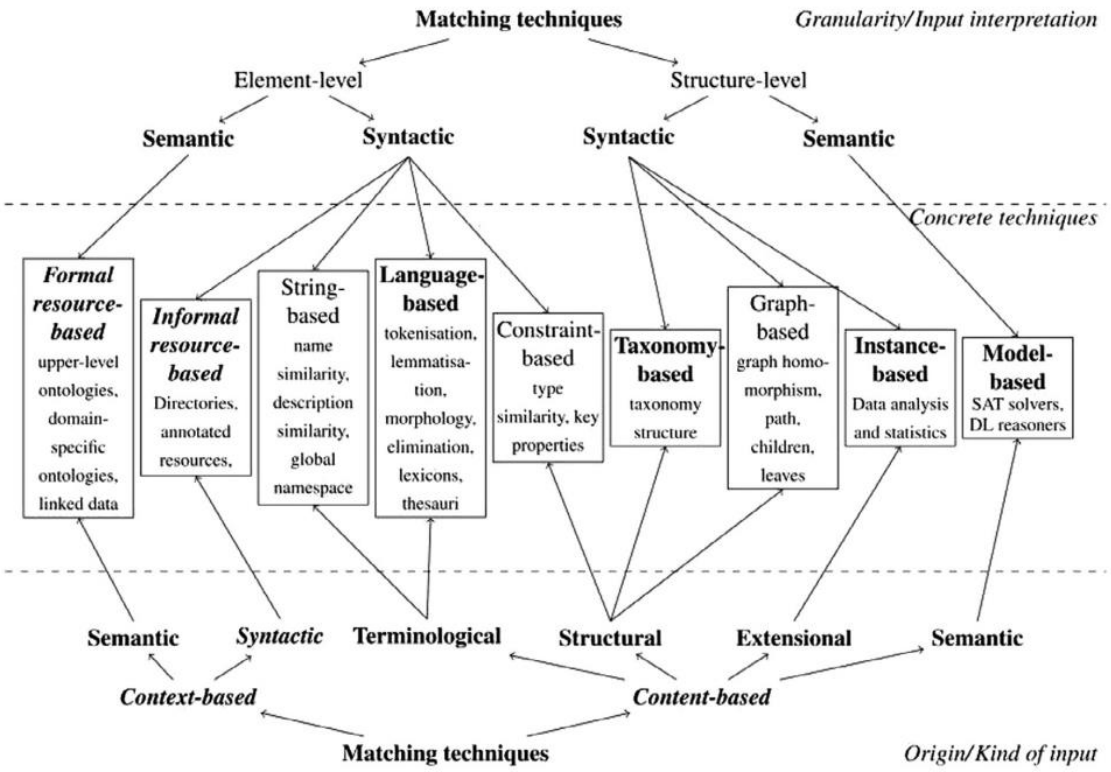


Note. Adapted from (Zeng 2019). Mapping models recommended by ISO 25964 (based on ISO 25964-2:2013, 6.3 and 6.4)

Euzenat and Shavaiko (2013), in their study, present a classification of matching approaches that also summarizes the approaches taken to mapping, especially for ontologies. They indicate that these matching techniques might be employed at the semantic or syntactic level, as outlined in Figure 5.

Figure 5

Ontology Matching Techniques



Note. Adapted from (Euzenat & Shvaiko, 2013). Classification of matching approaches.

In this classification, the top layer focuses on granularity and input interpretation. The middle layer outlines various types of matching techniques. The bottom layer focuses on the origin of the information being used and the type of input that can be accepted.

Other factors can impact the mapping process results beyond the issues faced with selecting the right mapping methods. Challenges can occur when selecting a target or intermediary terminology due to issues with coverage and granularity of the terminology, continuous rapid updates of terminologies, and the time needed to evaluate them. Another factor that can affect the problem is the complexity of the terminologies, especially as they become more linked to other schemes.

Saitwal et al. (2012) state that time, effort, and terminology-specific expertise is needed to take on this challenging task. They noted that while granularity issues must be addressed, they are not always obvious, necessitating an extension of the instance hierarchy to preserve it. While they recognize that mandating terminology standards for clinical information systems will reduce the burden in terms of the number of mappings that must be made, they also recommend that better-automated methods for automatically linking concepts across schemas be developed. They further suggest that whatever the method is applied, it should be one that can be easily reproduced when source terminologies are updated (Saitwal et al., 2012).

2.2 Studies Outlining Mapping Approaches

Researchers have been testing automated approaches for a long time with varying levels of success. For example, Barrows, Cimino, and Clayton (1994) attempted to map clinically used diagnostic terms from a legacy ambulatory system to another controlled vocabulary in their clinical information system. Their methodology used lexical and morphological text matching algorithms to identify matches and verified by clinicians after that. These methods allowed for data to be migrated easily despite their differences and support patient care activities, yet these methods were still young when applied to this problem. The authors acknowledged that string matching and semantic indexing algorithms might outperform those they had tried.

Saitwal et al. (2012) have outlined the challenges they encountered mapping medication terminological systems. They attempted to map medications from a commercial electronic health record to a drug classification system to let researchers retrieve patient records from a clinical data warehouse based on indications or classes of prescribed medications. Mapping source codes to drug concepts such as ingredient, dose, and drug strength the researchers chose a concept that used as many of the medication codes as possible, when exact matches

were not available. Several methods for mapping including some automatic mapping methods that demonstrated the utility of string-matching algorithms were tested.

String-based methods were used to find the best SNOMED CT drug name for generics in the EHR and to match trade names to generic names using RxNorm mapped to SNOMED CT. Manual mappings using the SNOMED CT Browser to manually search for the drug names from the EHR, narrowing and then selecting the best matching concept were also tested. Mappings were evaluated according to the criteria of completeness, correctness, and accuracy. A human expert performed validation by reviewing a sample of drug mappings from each automated method and found that 45% of their mappings could be verified. The authors concluded that difficulties with correctly using and updating complex, rapidly evolving terminologies, difference in granularity and the time and effort needed to complete the mappings were challenges that remained to be addressed.

Natural Language Processing (NLP) is sometimes used to facilitate automated mapping between terminologies. Zhou et al. (2012) used NLP to create mappings at the term and concept level between RxNorm and local medication terminologies for interoperability and meaningful use requirements. Their MTerms tool comprises algorithms that perform exact match, data cleaning, re-sequencing, normalization, and conversion rules. In this project, the authors utilized measures that are commonly used for evaluating these types of algorithms. The automated mappings were evaluated by two reviewers using a set of qualitative evaluation metrics to rate the quality of the matches. The match type metrics used were exact match and partial match, further broken down into broader partial match, narrower partial match, incomplete partial match, and missing. The statistical measures of precision, recall, F measure, and accuracy were calculated for the mappings.

Zhou et al. (2012) found that different levels of granularity between the terminologies impact the mappings, requiring that workarounds be found to identify the closest matching concepts. This study showed significant time and labor reductions combined with high precision

in mapping terms and concepts. They suggested that algorithms be created based on evaluation metrics that could lessen ambiguity. The authors also found that missing terms were a challenge that could not be solved with automated mapping. Missing terms were due to differences in representation in terms of inclusion or exclusion from the source. The authors identified as a gap in the research, the need for systems that can be easily reproduced when changes in the target and source vocabulary occur.

Dias, Alves, and Felipe (2014) also attempted mapping between terminologies in healthcare to achieve semantic interoperability. They intended to integrate two separate databases with the same information previously encoded using different terminologies. This rule-based approach used association rules mining for knowledge extraction, which represented translations between the terminologies. Domain experts then used the extracted rules and their measurements to determine whether the relationships obtained were an accurate translation for the terms or not. Whenever rules could not be obtained, they also used text search string matching between terms.

Results showed that extracted rules make it easier for experts to define correct mappings because the system will use those rules to suggest codes making it easier to map from those rules than from the results of the text search. This method relies on the expertise of human coders to select the correct mapping with the help of the automatically generated rules. In addition, many times, the experts rejected the suggestions made. Therefore, challenges remain with automatic matches that indicate a human expert must still verify the results. The research also suggests that the burden on the expert is reduced since rather than having to search the terminology manually they can simply use the suggestions reduce their workload.

Another study aimed at semantic interoperability in healthcare applications highlights the dependence on controlled terminologies to enable inter-machine exchange. The authors designed a framework that exploits mapping approaches for finding similarities between terminologies, uses experts to validate the mappings, and additionally uses a reasoner to

identify inferred mappings and to validate asserted and inferred mappings (Hussain et al., 2014). Their framework aims to provide provenance information with the mappings as they identify it as necessary to accompany contextual information. They have found that their framework enables a more collaborative semantic landscape and that it can provide usage data and feedback mechanisms for institutions that provide mappings.

Allones, Martinez, and Taboada (2014) used automated mapping techniques to codify procedures in pathology. Their solution identified text-to-concept mappings in SNOMED CT. It used name-based, terminological, structural, and disambiguation techniques to find text-to-concept mappings. Heuristic rules were created to aid with selecting more accurate mappings, and experiments were designed for evaluation which tested precision, recall, and the F-measure (a weighted average of precision and recall). The results demonstrate that query expansion helps improve recall and that disambiguation techniques yield excellent results. This tool uses two separate matching profiles, one is fully automatic, and one is semiautomatic. The results show that a fully automatic process makes it possible to achieve mapping without the need for expert oversight. However, the authors mention a need for a framework that can combine different techniques and be applied to various terminologies.

In another study that tested a variety of techniques for mapping (Kolyvakis et al., 2018). The researchers tested feature engineering which involves using features of the data that a machine-learning algorithm can utilize to achieve specific tasks such as mapping. They also tried binary classification which uses a classier trained on a set of alignments to make predictions of semantic equivalence between concepts. However, due to class imbalance issues stemming from the completely different data models, the results suffered (Kolyvakis et al., 2018). The authors suggested testing unsupervised learning methods e.g., neural representations to address these problems.

Kolyvakis et al. (2018) also used machine learning to map words from high dimensional vectors which consider the context. Words that appeared in a similar context had similarity

measures applied. The aim was to capture and exploit the context in which words are used in definitions and synonym relations to make inferences about the concepts. Pairwise and cosine distance calculations were used to make evaluations, and outlier detection was used to detect differences between semantic similarity and related terms. The authors found that the unique and rare words used because of the domain's specificity made deep learning challenging and impacted the matching task. Choosing the best similarity metrics is a complicated process that requires tuning of thresholds used in these metrics. They found that ontology matching can be performed without structural information but that there still needs to be a determination about how structural information can best be exploited for mapping.

2.3 KOS Tools for Mapping

Liang et al. (2016) designed MeTMapS to address the limitations of other terminology systems that required prior knowledge about the mappings, making it complicated to load terminologies. When versions changed, mappings had to be recreated and reloaded. Their APIs did not conform to standard specifications, and differences in schema made transforming and visualizing data difficult. Browsing and filtering were found not sufficient for efficient searching. Their system aimed to address those issues by reusing the best features of these terminology systems but addressing their limitations. Their solution was designed as a web application with relevant information displayed on a single page to facilitate navigation and information processing. The search results are organized into a tree structure. The study results showed that their systems were able to show correct terms first, with the most relevant being shown at the top. The system could suggest terms while typing, handle exact term matches for different concepts, suggest generic descriptive names, and select multiple terminologies for mapping. However, the system could not handle misspelled terms and partial words and needed to expand its search functions to handle known synonyms.

YAM++ was developed for ontology and thesaurus matching through a map validation and enrichment interface. It proposes a solution to mapping that allows domain experts with basic technical knowledge to accomplish mapping and alignment and validation of the matching tasks (Bellahsene et al., 2017). It uses a variety of matching algorithms to discover equivalence relationships between ontology elements. In particular, it uses a terminology matcher that produces mapping that compares annotations, an instance-based matcher that supports new mapping based on shared instances, and a contextual matcher that computes similarity values between entities. It then performs post-filtering and semantic verification to select and check the consistency of mappings. Requests for mappings are submitted through an HTTP API, and a validator module allows a manual expert to validate the automatically generated mappings. The authors see a need for alignment validation through crowdsourcing.

The Unified Medical Language System and BioPortal both support interoperability through the integration of multiple vocabularies. The UMLS is often used in applications that enhance access to medical literature (Bodenreider, 2004). BioPortal is a repository for biomedical ontologies that supports analysis, visualization, and download of large datasets. With BioPortal, anyone can submit their ontology and mappings, and there are few constraints on the ontology submissions beyond being related to biomedicine and using an appropriate format (Noy et al., 2009; Salvadores et al., 2013). This lack of requirements means that there will be many differences between ontologies in size, quality, expressivity, and complexity. However, BioPortal does offer more opportunities for visualizing data in ways that are not available in UMLS. Both however, can be used to facilitate the development of systems that use natural language processing to create medical informatics solutions for researchers, and both provide a web interface that can be searched and browsed to explore terms and relations between terms. The UMLS and its associated MetaMap tool is often used in mapping projects.

2.4 Mapping and Semantic Analysis

Semantic analysis aims to minimize syntactic structures and provide meaning, find synonyms, word sense disambiguation, translate from one natural language to another, extract entities and relations, and populate a base of knowledge. Knowing the semantic correspondences between their elements is essential to address semantic mappings among disparate coding schemes. Thus, semantic analysis provides methods and models for extracting pertinent information from unstructured data. If relationships exist between important entities in the document, these relationships can be represented, supporting reasoning and inferencing and creating new knowledge (Davies et al., 2006).

In a study by Zhu et al. (2013), text mining techniques were used to extract novel knowledge from scientific text. The authors mention that this method can employ many semantic technologies, including machine learning, natural language processing, and pattern recognition, to find hidden outcomes in unstructured text. They identify named entity recognition as the most important step in extracting knowledge since it identifies terms or concepts. It can be performed through dictionary-based approaches, which can miss undefined terms that are not mentioned in the dictionary. Alternatively, NER can be performed through rule-based approaches to identify terms from text though not always effective, or machine learning approaches that use standard annotated training data sets are data-driven and application domain-oriented. They suggest using precision, recall rate, and F_1 (accuracy) rate to evaluate the performance of this approach.

Chen et al. (2020) assessed trends and specific attributes of natural language processing techniques used for clinical trials text analysis in a more recent study. They found NLP to be an effective tool for obtaining structured information from unstructured data. Their study notes that a significant portion of clinical trial information is documented and stored within the unstructured text portions and that unlocking the hidden knowledge and enabling advanced reasoning can be accomplished by adopting NLP techniques.

The authors found that text mining and artificial intelligence approaches were most often used. They also found that hybrid approaches had much success and were being more commonly utilized. An example would be combining pattern-driven, knowledge-enriched, and feature-weighted approaches. Deep learning involving neural networks was also another more commonly used method. In addition, rule-driven frameworks that combine lexical, syntactic, and meta-level, task-specific knowledge inputs were also useful.

Colic, Furrer, and Rinaldi (2020) also used natural language processing to perform named entity recognition (NER) and text summarization of COVID-19 data. Their study focused on identifying relevant scientific literature by identifying terms through NER and mapping them to unique IDs in a controlled vocabulary. Their approach combined entity recognition and linking by simultaneously identifying interest entities and mapping them to appropriate entries in the various coding schemes. This study used clinical coding schemes such as including Chemical Entities of Biological Interest (CHEBI), NCBI Taxonomy (NCBITaxon) and is the only one found that specifically includes a scheme (COVOC) focused on COVID-19.

The techniques used in these studies to understand text are many and varied and include parsing, stemming, text segmentation, named entity recognition, relationship extraction, sentiment analysis, and deep learning. With the urgent needs researchers face to find solutions for dealing with the pandemic, semantic analysis offers methods for improving the way information is presented. It supports merging information from all relevant documents, removing redundant information, and summarizing portions of the information.

2.5 Assessment of Studies, Gaps, and Justification

These studies and tools suggest a need for a simple approach to mapping that is easily leveraged and replicated yet does not require human oversight. Various researchers have outlined issues they feel should be addressed in mapping tasks (Arvanitis, 2014; M. Zeng, 2019), specifically for considering syntax and semantics. However, most studies do not consider

both issues in their solutions. The focus is instead placed on leveraging specific algorithms for specific use cases or institutions. Therefore, aside from the openness of KOS tools listed (Bellahsene et al., 2017; Bodenreider, 2004; Liang et al., 2016; Noy et al., 2009; Salvadores et al., 2013), the solutions created can only be accessed in the ways they are made available. There is limited opportunity for a user to reuse or re-engineer the solution to meet their needs. In addition, there are time and labor constraints mentioned in some studies (Saitwal et al., 2012; Zhou et al., 2012) that could make it hard for others to reuse them. Finally, some of the studies (Allones et al., 2014; Bellahsene et al., 2017) recommended that various approaches be implemented to solve the problem rather than using one or two solutions.

This research study offers a unique perspective on the issue of interoperability and mapping. It seeks to develop a tool that focuses on a new problem area, COVID-19, through designing and developing a novel tool that combines mapping and annotation. Other studies have been more broadly focused on specific biomedical applications or directly mapping terminologies. While mapping tools have been built as outlined in the section on KOS tools for mapping, they were created before COVID-19 and are not specific to any one context. Also, no solution was found that attempted to combine mapping and annotation of clinical documentation. Therefore, a tool that can target these new vocabularies and integrate them with those already being mapped will be useful, if not critical.

The studies reviewed showed the utility of NLP applications and the implementation of algorithms, machine learning, and semantic web techniques for mapping and annotating unstructured text. The results of several of the studies seem to suggest that approaches that combine approaches in a way that answers the challenge at hand are better than using only one specific approach. In this research, similar techniques will be applied inside a workflow that will reduce the need for programming knowledge and mapping expertise. It will be replicable in new contexts and easily deployed. None of the studies have attempted to combine both mappings of clinical coding schemes generally and annotation of text in a single tool. Also, issues related to

mapping terminologies with different levels of granularity and specificity have still not been solved, and automated methods were not able to handle the problem of missing terms. It may be possible to explore whether these issues show up in the workflow design process or any solutions to that problem. This study will demonstrate that a novel workflow designed with ETL tools combining mapping methods might address some of the issues faced by these projects and provide a simpler, more easily adaptable mapping method.

Chapter 3. Methodology

3.1 The Design Science Research (DSR) Approach

This section describes the structure of the research design and outlines the methodology used to conduct this research study. The workflow developed explores the mapping of clinical coding schemes using the KNIME platform. Vaishnavi and Keuchler (2015) note that research in information systems, science, and communication technologies is multi-paradigmatic. The research questions, methodologies, and grounding philosophies are drawn from multiple fields and united under common interests that seek to understand how human-computer systems develop, produce, and process information and impact the organizations in which they are embedded.

There are multiple ways in which research might be undertaken and researchers should be aware of the choices made during the research process and the potential impact of those choices. A DSR approach guiding the development of artifacts as objects of research will be used to address the following questions:

Research Question 1. How can an Extract Transform Load (ETL) workflow tool support the task of clinical coding scheme mapping?

Research Question 2. How does the mapping output of the novel workflow support and affect annotation of clinical trials in COVID-19 research?

Research Question 3. How can the sociotechnical model be leveraged or updated to explain and assess mapping to achieve semantic interoperability in clinical coding schemes?

3.2 Methodological Grounding

Design Science Research (DSR) has been chosen as the philosophical and methodological approach to support the discovery and identification of opportunities and problems relevant to clinical coding scheme mapping and the development of a workflow to

address that issue. The seminal paper on DSR in Information Systems research by Hevner et al. (2004) outlined a framework for guiding research with the artefact as the main goal and is ideal for creating and evaluating artifacts that will be used to solve identified problems. Hevner and Chaterjee's (2010b) definition of design science research states that it is:

> *"A research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem."*

Simon (1996), in his book The Sciences of the Artificial, states that design science differs from natural science with its emphasis on "knowledge about some class of things – objects or phenomena – in the world (nature or society) that describes and explains how they behave and interact with each other." DSR is instead concerned with "knowledge about artificial (man-made) objects and phenomena designed to meet certain desired goals" (Simon, 1996). DSR evaluation is also different from natural science or theory driven behavioral science experimentation in that iteration is critical between design (development) and evaluation (experiment) (Kuechler & Vaishnavi, 2008). In natural science, the experimental procedure, apparatus et cetera are designed to minimize confounding factors and clearly support or disconfirm theory. However, in DSR “both the artifact and the experimental setting are intentionally complex (and thus confounded) in order to develop methods and artifacts that are useful in practice” (Kuechler & Vaishnavi, 2008).

The output of Design Science Research should take the form of a knowledge contribution in the form of either an invention, improvement, or adaptation. That is, the researcher should either invent new knowledge or solution for a new problem, develop new knowledge or solution for a known problem, or adapt a known knowledge or solution to a new problem. It is also possible to make more than one kind of contribution in a single research

project (Gregor & Hevner, 2013). These knowledge contributions or artifacts can more specifically be constructs, models, methods, instantiations, frameworks, social innovations, new properties of technical, social, or informational resources, and design theories (Vaishnavi et al., 2015). Artifacts include any designed object with an embedded solution to an understood research problem. Therefore, the objective of creating and testing a novel workflow using a particular tool to solve the real-world problem of mapping clinical coding schemes, aligns with the tenets of design science research.

Various models of the DSR research process have been presented. Table 1 presents a review of these. Based on Table 1, we can infer that the most critical features involved in DSR are a) understanding the problem, b) development and evaluation of the artifact, and finally, c) communicating gained knowledge.

Table 1

Comparison of DSR Research Methodology Steps

Lukka (2003)	**Vaishnavi and Kuechler (2015)**	**Peffers et al (2008)**	**Kasanen et al (1993)**
Identify a practically relevant problem with theoretical contribution potential	Awareness of problem	Identify a problem and motivate	Find a practically relevant problem that has research potential
Examine the potential for long-term research		Define objectives of a solution	

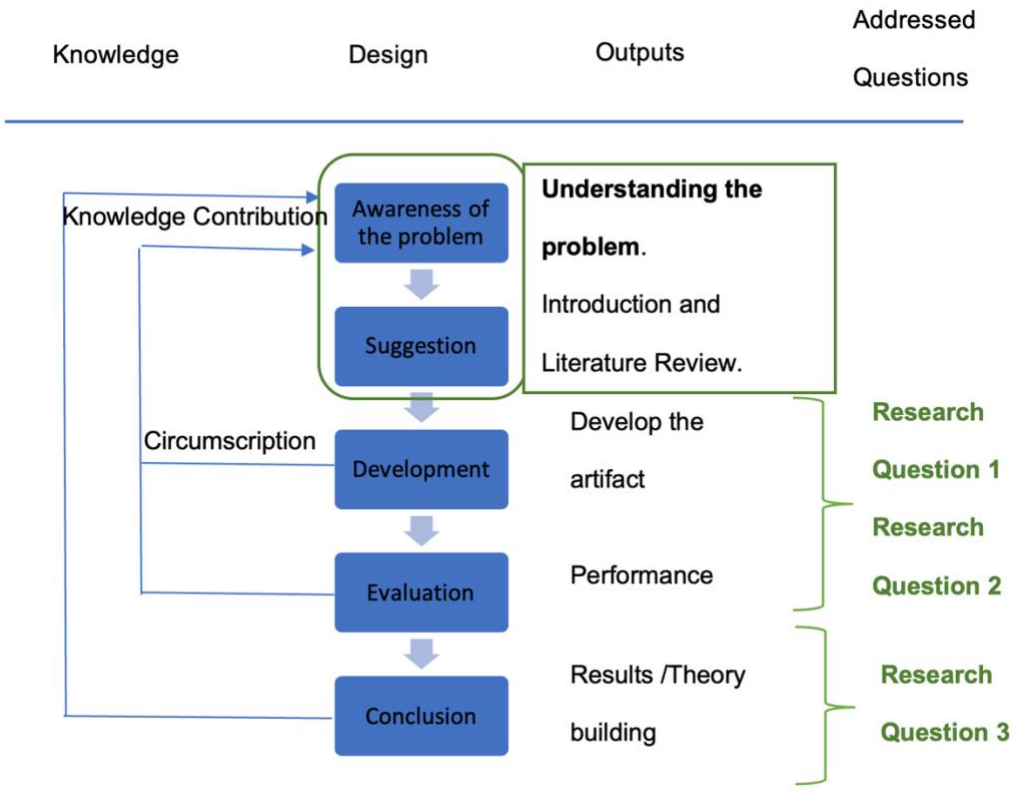
cooperation with target organizations			
Gain a deep understanding of the topic area in theory and practice			Obtain a general and comprehensive understanding of the topic
Have ideas for the solution and develop a problem-solving construction	Suggestion Development	Design and development	Innovate, i.e., construct a solution idea
Implement the solution and test how it works	Evaluation	Demonstration Evaluation	Demonstrate that the solution works
Consider the scope of applicability of the solution			Examine the scope of applicability of the solution
Identify and analyze the theoretical contribution	Conclusion	Communication	Show the theoretical connections and the research contribution of the solution concept

This study models the research design to match the cognitive processes involved in the DSR cycle as outlined in Figure 6. The model is adapted to reflect the three main critical

features identified from the review of models of the research process above and shows the areas in which research questions will be addressed. A brief discussion of these critical steps and how they apply to this study are addressed in section 3.3.

Figure 6

Research Plan Aligned with the Cognitive Model of DSR



Note. Adapted from (Vaishnavi et al., 2015), Cognition in the design science research cycle.

3.3 Research Implementation Plan

3.3.1 Understanding the Problem

The first tasks of the DSR process are identifying the problem in the applicable environment, considering the existing knowledge, and understanding the research

requirements. The researcher describes and explains the problem and discusses the remaining issues if nothing is done to address the problem (Wieringa, 2009). Problems can be investigated from several categorical perspectives. A description of these categories, according to Wieringa (2009), follows:

- The problem-driven investigation begins with an identified problem and a diagnosis of the problem and its causes to determine how it should be solved.
- The goal-driven investigation does not necessarily involve a problem at all. Instead, it starts with an analysis of goals that need to be achieved and develops a plan for achieving them.
- The solution-driven investigation starts with investigating the properties of a technology and explores ways in which it can be used to solve a problem or achieve a goal.
- And the impact-driven investigation uses the outcomes of past actions rather than designing future solutions. It researches and describes solutions implemented, examines their impacts, and translates these into criteria to be applied in a new context.

In previous sections, a description of the function of clinical coding schemes in helping to achieve meaningful and accurate information exchange, knowledge enrichment, and data analysis is given.

Challenges to semantic interoperability, health information exchange, analytics, and research posed when there are no mappings across schemes are also highlighted. The difficulties faced due to manual, time-consuming, error-prone, or overly complex mapping methods are further outlined. Therefore, presenting a solution to these problems is one main objective of this research study. The solution-driven design where the potential of a particular technology to solve a problem is examined fits this research study best. However, some aspects of the study are problem-driven based on identified problems in the domain being addressed. Wieringa (2009)notes that the categories are not mutually exclusive and can be present in

multiple, however depending on the issues at hand, some aspects will be emphasized, and others de-emphasized. Vaishnavi and Kuechler (2015) refer to this as the Preliminaries Type Pattern of DSR Research Design and Development, where tools and techniques applicable to the problem space are identified and used to obtain knowledge relevant to the research question and demonstrate worthwhile techniques that have not yet been used in the problem space.

Two small pilot studies completed prior to this work informed the choice of tool and methods used in this work. A pilot study is a small-scale study that asks whether something can be done and, if so, how. It assists with the planning and modification of the main study. They can be performed either externally, i.e., independent of the main study, or as part of the research design of the main study (In, 2017). Automated mapping approaches were tested in a related pilot study focused on coding schemes in the library science domain. Another small study evaluated NER on clinical trial data to inform the processes that would need to be implemented to achieve it in the workflow tool. These studies are described in the following sections.

3.3.1.1 Pilot Study 1 – Resources and Vocabulary Enrichment for Analytics. This project used data from the Digital Public Library of America, which is an open distributed network of comprehensive online resources that aggregates data from over 42 hubs across the nation, consisting of data from libraries, universities, archives, historical societies, and museums. The project's tasks were to explore, develop and test effective methods to analyze record content and match content, including keywords, with the Library of Congress Subject Headings (LCSH) and the Art and Architecture Thesaurus (AAT).

A snapshot of their entire collections of 22,158,160 items (refers to resources belonging to collections, e.g., images), showed that a majority of them (n=10,698,050) had one to six subject headings. The data value standards that were most commonly aligned with those subject headings include the Faceted Application of Subject Terminology (FAST), Medical

Subject Headings (MeSH), AAT and LCSH, among others. However, there were also many items with no subject headings.

Since data analytics was a big part of the project, several tools in the data science space were considered and Alteryx Data Analytics Platform chosen. Alteryx allows easy manipulation of data without writing complicated codes through a workbench approach where workflows are built from a menu of nodes. The LCSH and AAT were obtained as N-triple serialization and converted to CSV format from which URIs, and subject terms were extracted. To reduce computing costs, a sample of the dataset was used in the created workflow. The workflow connected nodes (icons which encapsulate functions) for data cleaning and transformation, joins and unions to find exactly matching terms and fuzzy match nodes which implemented string similarity algorithms to find similar terms.

The results for a sample set of 500,000 terms were modest. From the workflow 21027 exactly matching terms and an additional 323263 unique fuzzy matches in the second stage, representing close, and partial matches to terms in LCSH were identified. In the AAT match workflow 6898 exactly matching terms, 26, 236 fuzzy matches representing close and partial matches were found. The results showed that exact and close matches could be accepted without human review but that partial matches would need to be checked by a domain expert for labeling with the appropriate semantic type relationships.

Due to a short study timeframe, further refinement and development were not possible, though testing of ML approaches were thought to be an ideal next step. The study was also limited based on certain assumptions made about the dataset such as that the terms assigned to the items were also in alignment with a controlled vocabulary. In practice, this was not always true. In addition, there was not enough computing power necessary to quickly process large amounts of data and used a sample to reduce the complexities. For some nodes and functions in Alteryx, processing is both time and resource consuming. This brought new challenges since

it was not possible to select only items which used terms aligned with the aligned to the tested controlled vocabularies.

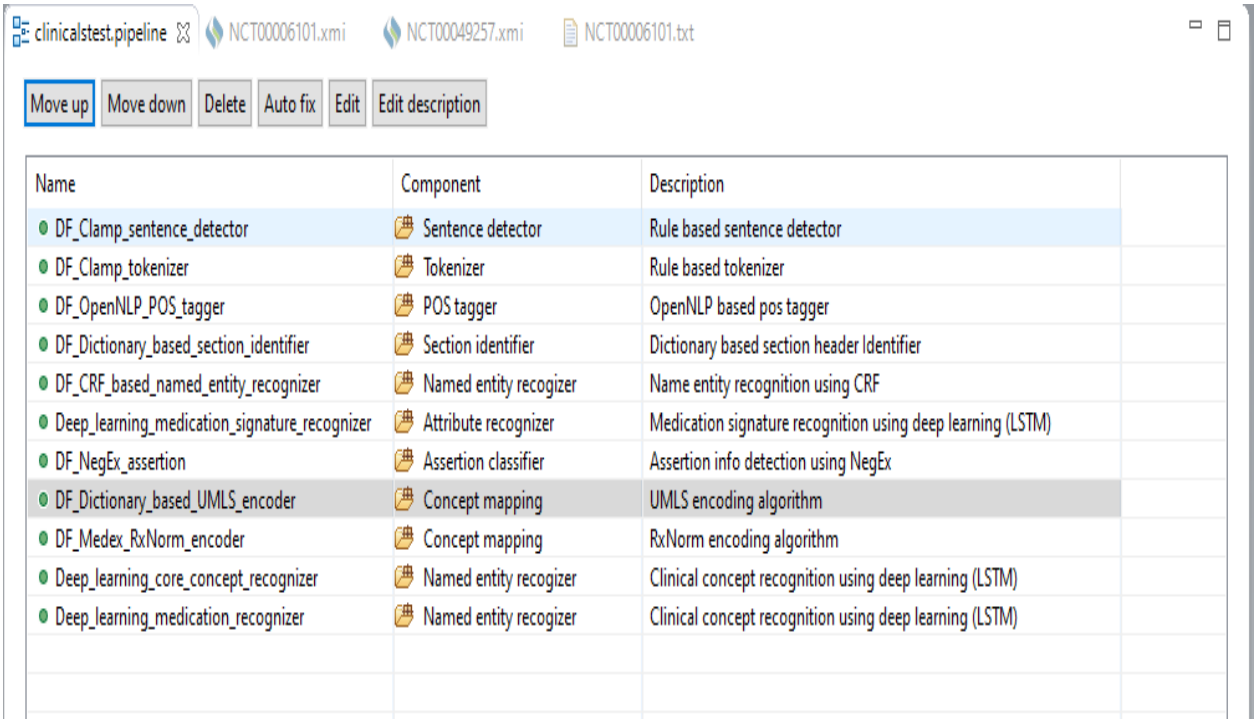
3.3.1.2 Implications of Pilot Study 1. This pilot informs certain aspects of the current study. Most importantly, the process and results suggested that Extract, Transform, Load (ETL) tools like Alteryx, had some utility not yet fully explored for mapping between controlled vocabularies. Mapping is time consuming, resource intensive, and usually requires expert knowledge to complete. The pilot study sparked an interest in testing to what extent this kind of tool might be used in the problem space and what knowledge attempting it would yield.

Since in the pilot, controlled vocabulary terms were used for enrichment it was hypothesized that any scheme from which term labels could be extracted, could be manipulated in a similar way. For the workflow created in this work, the methods for data input were informed by the pilot. Particularly, what data elements should ideally be extracted for vocabulary enrichment and how those elements are commonly found in RDF datasets. Finally, the process of lexical matching in this work is guided by the process and lessons learned in the pilot. For example, choosing algorithms, setting thresholds for what can be considered matching terms, or identifying exact matches.

3.3.1.3 Pilot Study 2 – Annotation testing in CLAMP*.* Tests were performed with the tool CLAMP, where a natural language processing pipeline was implemented. The pipeline took as input clinical trials in text format, encoded with the Unified Medical Language System (UMLS) and RxNorm vocabularies. These were selected because they had the most coverage of medical and pharmaceutical terms. The NLP pipeline in Figure 7 used a sentence detector, tokenizer, part of speech tagger, named entity recognizer, and concept recognizer to parse the text before the encoders for UMLS and RxNorm were applied. A limitation of this pilot is the use of the default training data instead of creating a training set specific to a certain context.

Figure 7

A Pipeline of Natural Language Processes for Extracting Data from Clinical Trials



The results from tagging are shown in Figures 8 with semantic types. The data is tagged according to whether the identified entity was a problem, treatment, or drug and had a concept unique identifier (CUI) assigned. Results showed the pipeline assigned appropriate semantic types and CUIs to the extracted entities; however, some misidentified entities and entities with no returned CUIs were also present. An explanation for missing entities is that the terms used in the clinical trials are not controlled or do not align with the preferred or alternate terms in the encoding vocabularies present within the UMLS. Missing entities could be addressed by adding other vocabularies into the pipeline for parsing the data, more specific to the domain in question.

3.3.1.4 Implications of Pilot Study 2. The pilot with CLAMP assessed using an NLP pipeline method for the extraction and annotation of clinical trial data. It was helpful for informing the necessary NLP processes and the order of application that would be ideal for accomplishing entity recognition and tagging in the workflow tool. In addition, it provided an example of the kind of output that should be produced by the annotation portion of the workflow tool. The results

suggested that data could be more efficiently tagged if the NLP model was to be trained with COVID-19 data and tagged with controlled vocabulary terms specific to COVID-19.

Figure 8

Extracted and Encoded Clinical Trial Results

A	B	C	D	E	F	G	H
Start	End	Semantic	CUI	Assertion	Entity		
219	223	test	C4196219	present	UCLA		
219	223	test	null	null	UCLA		
302	313	treatment	C0728747	present	Trastuzumab		
302	313	drug	null	null	Trastuzumab		
318	327	drug	null	null	Erlotinib		
318	327	problem	null	null	Erlotinib		
331	349	treatment	null	null	First-Line Therapy		
337	349	treatment	C1708063	present	Line Therapy		
373	397	problem	C0278488	present	Metastatic Breast Cancer		
373	397	problem	null	null	Metastatic Breast Cancer		
414	418	problem	C0069515	present	HER2		
463	472	treatment	C0338204	present	Herceptin		
463	472	drug	null	null	Herceptin		
487	494	treatment	C1135137	present	OSI-774		
487	490	test	null	null	OSI		
508	522	treatment	C1708063	present	Line Treatment		
526	550	problem	C0278488	present	Metastatic Breast Cancer		
526	550	problem	null	null	Metastatic Breast Cancer		
567	571	problem	C0069515	present	HER2		
600	628	problem	C1517622	present	Jonsson Comprehensive Cancer		
660	675	problem	C1513882	present	National Cancer		
753	781	problem	C1517622	present	Jonsson Comprehensive Cancer		
826	847	test	C0003250	present	Monoclonal antibodies		
826	847	test	null	null	Monoclonal antibodies		
856	867	treatment	C0728747	present	trastuzumab		
856	867	problem	null	null	trastuzumab		
872	884	problem	C0475446	present	locate tumor		
879	890	problem	null	null	tumor cells		
922	935	problem	C0157013	present	deliver tumor		

Start	End	Semantic	CUI	Assertion	Entity
7645	7661	problem	null	null	motor neuropathy
7692	7722	problem	null	null	known carcinomatous meningitis
7698	7722	problem	C0220654	present	carcinomatous meningitis
7726	7742	problem	C0220650	present	brain metastases
7726	7742	problem	null	null	brain metastases
7815	7823	problem	C0027651	present	neoplasm
7815	7823	problem	null	null	neoplasm
7876	7900	problem	C0347073	present	in-situ carcinoma of any
7907	7931	problem	C0699893	present	non-melanoma skin cancer
7907	7931	problem	null	null	non-melanoma skin cancer
7936	7952	problem	C0155619	present	other malignancy
7936	7952	problem	null	null	other malignancy
7978	7985	treatment	C0543467	present	surgery
7978	7985	treatment	null	null	surgery
7989	7998	problem	C0332301	present	radiation
7989	7998	treatment	null	null	radiation
8004	8013	problem	C0086651	present	a disease
8004	8013	problem	null	null	a disease
8092	8105	treatment	C0679637	present	major surgery
8092	8105	treatment	null	null	major surgery
8164	8170	test	C0005558	present	biopsy
8164	8170	test	null	null	biopsy
8187	8209	treatment	null	null	a venous access device
8189	8209	treatment	C0262741	present	venous access device
8280	8297	treatment	null	null	any prior surgery
8284	8297	treatment	C0455610	present	prior surgery
8341	8364	problem	C0455684	present	serious cardiac disease

3.3.2 Development of the Artifact

The pilot studies both clarify and inform important workflow goals within the project. Choosing appropriate algorithms and rule-based approaches for matching terms and using the integrated mapped term set to create dictionary that can be used for tagging clinical trial data using NLP. The pilot also demonstrated the utility of using a certain type of technology, but answers are still needed about how helpful and efficient this process might be, how it impacts annotation of clinical trials and what knowledge can be gained from using it. Therefore, in this work, a novel workflow using KNIME to integrate clinical coding schemes and annotate clinical research data is developed. Findings and lessons learned are used to enhance knowledge by answering the research questions.

The design and development phase of the DSR process covers artifact development. Generally, techniques may vary depending on the aims of the artifact (Vaishnavi et al., 2015). Lukka (2003) describes this stage as fundamentally creative and exploratory by nature, not conforming strictly to a particular methodology. Vaishnavi et al. (2015) describe it as testing the methods from the suggestion or literature review for accomplishing the solution. They further describe the iterative nature of a process where problems are imperfectly understood and multifaceted, requiring exploration and experimental methods to solve as well as backtracking to reassess if one solution impedes another.

Before creating the workflow, the structure and properties of the selected terminologies, COVID-19 Vocabulary Ontology, COVID-19 ontology and the Coronavirus Infectious Disease Ontology were assessed, followed by selection and transformation of the data. These vocabularies are selected because they are directly created for or related to COVID-19 research. The outcome of studies in the literature review suggested that combining methods often produces better results than a single method (Allones et al., 2014; Bellahsene et al., 2017; Colic et al., 2020; Hussain et al., 2014). Therefore, within the workflow, a variety of methods including string-based matching algorithms that have demonstrated efficiency for locating

occurrences of terms and specific patterns of text and matching and clustering entity names (Aho & Corasick, 1975; Cohen et al., 2003; Monge & Elkan, 1997); sense-based algorithms that function by making matches based on relations (Giunchiglia et al., 2004); and rule-based methods that determine the semantic type of matches are tested. Focusing on the semantic rather than the syntactic reduces incorrect mappings (M. Zeng, 2019), therefore combining those kinds of methods should improve semantic interoperability outcomes.

Mapping degrees were assessed based on properties from RDFS, OWL, and SKOS. Zeng (2019) describes these as mappings between ontological classes, between properties, concepts from concept schemes, or transitive super properties. With the focus placed on concepts, SKOS labels for the mappings can be applied as skos:exactMatch, skos:closeMatch, skos:relatedMatch, and so on. Once all terms were mapped, an integrated set of concepts were used to create a dictionary for the NLP pipeline. The annotation of clinical trials for COVID-19 was the final task to be implemented within the novel workflow. The methods used within this workflow are also informed by research and consist of standard NLP tasks, such as entity extraction, tagging, and bag of words. A conceptual model of the design process was presented in Figure 3.

3.3.2.1 Design Tool. KNIME is an open-source ETL (extract, transform, load) software for data loading, transformation, analysis, and visual exploration. It features a graphical user interface where a researcher can link together blocks representing steps in a data science workflow. In KNIME, it is possible to perform ETL processes, machine learning, deep learning, natural language processing, API integrations, statistical inference, and interactive visual analytics. This integrated development platform also allows customization through an extensible plugin system, so researchers can build in features they need using Python, Java, R, Scala, or use community plugins.

KNIME is an ideal tool for solution-driven design processes where a researcher can create workflows for users who are not traditionally trained programmers. Further, the

documentation tools built into the platform's feature set, its support for task assessment, and the grouping of tasks makes it easy to follow a workflow and understand what is happening visually or reproduce it with similar data and new contexts. From the pilot study it was determined that ETL tools could be helpful in the problem space of terminology mapping. Therefore, the KNIME data analytics platform was chosen to explore NLP, machine learning, and semantic analysis techniques for clinical coding scheme mapping and clinical trial annotation.

The pilot also highlighted the necessity of choosing a tool that could handle larger data sets with less computing resource requirements, which would reduce the hardware expenses of a project. Second, where Alteryx is commercially licensed, KNIME is open-source, which allows not only for ease of access and implementation, but gives access to the many plugins which have been created to add functionality to the tool. This meant that work done using this tool could be easily transferred to new contexts without high licensing cost. For comparison, the workflow from the pilot study cannot be used as is unless the new user obtains a license for the software. That means only the results from the processes, for example csv files, and the methods as knowledge are transferable. Further, KNIME's support for modification and collaboration, makes it easy to adapt the workflow with minimal complexity while other tools require technical knowledge and expertise for use and deployment on a local machine, or are constrained to one specialized task.

3.3.2.2 Sample Data Collection for Clinical Trial Annotation Workflow. This research used secondary data obtained from www.clinicaltrials.gov. This website is a registry of clinical trials made available by the National Library of Medicine. The National Institutes of Health define a clinical trial as "a research study in which human participants are prospectively assigned to one or more interventions to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes" (National Institutes of Health, 2017). This resource was chosen for ease of access. Other datasets had requirements for obtaining special

permissions, or the need to compete for access with a research proposal or use the dataset within a proprietary analytics environment.

Since this is a new disease, much work is being done in this area, testing old drugs in new contexts, examining impacts of COVID-19 on other conditions, et cetera. The applicability of the use of datasets from this source seemed justified based on these factors. The author will select clinical trials within the parameters shown in Table 2. The research used completed interventional studies with male and female adults over 18 years old. Study requirements for children are more stringent so to avoid ethical and reduce logistical challenges, this research only uses secondary study data for adults. Additionally, upon review, the number of studies available in this registry currently involving children is negligible therefore they are excluded.

Table 2

Clinical Trial Inclusion and Exclusion Criteria

Clinical Trials	
Inclusion Criteria	Interventional Adults (18-64) Older Adult (65+) Completed
Exclusion Criteria	Child (birth-17) Observational Studies
Keywords	COVID-19, SARS-CoV-2

The search terms for identifying these studies included COVID-19, coronavirus disease, and SARS-CoV-2. An interventional study is defined as a type of clinical study that assigns participants to an intervention or treatment group so researchers can evaluate the effects of the

intervention on health outcomes. Participants might get either diagnostic, therapeutic, or other types of interventions. Interventions can include drugs, medical devices, vaccines, and other procedures, products, or changes to behavior that induce some change (National Library of Medicine, 2021).

873 studies meeting these criteria were downloaded from the clinicaltrials.gov registry on November 21, 2021. From that set a sample of studies large enough to achieve a good estimate of model performance was selected. Determining an appropriate sample size of required text documents can be calculated from the corpus of terms that will be used (Figueroa et al., 2012; Juckett, 2012). In general, a recommendation of about 500 sample documents for a 95% capture probability has been found to be sufficient for most scenarios. Juckett (2012) specifically suggests using 80 - 560 sample documents with higher numbers yielding better results. Therefore, in terms of attrition, even if some proportions of the samples are unusable, if the number is between the recommended values, it should be possible to obtain acceptable results. Clinical trial data is made available from the website in portable document format (pdf), plain text (txt), tab and comma-separated values (TSV/CSV), and extensible markup language (XML). The full study records are available in XML format only.

Those XML records were downloaded and converted to comma separated values (csv) before being connected to the workflow. The connected csv data table had each clinical trial occupying a single row. To gain the sample corpus, the row sampling node was configured to select 90% of the rows (i.e., clinical trials) in a random sampling of all rows, other options include linear and stratified sampling, or to select the top rows of data. A random seed was used to ensure reproducible results. This resulted in a set of 785 terms, from which we removed documents with missing data elements such as missing descriptions leaving a total of 575 documents which is just above the higher end of the recommended range. Missing descriptions are an unstructured text fields describing the features of the study, the plan, methods, et cetera.

3.3.2.3 Clinical Coding Scheme Access. The clinical coding schemes used in this study were created for naming concepts related to COVID-19, such as symptoms, contact tracing, infection rates, and drug testing. Specifically, the Coronavirus Vocabulary Ontology (COVOC), COVID 19 Ontology, and the Coronavirus Infectious Disease Ontology (CIDO) were used for concept mapping. LOINC was also selected for inclusion, particularly as a means of testing the mapping results. These were obtained in their full triple format from the repository where they are stored and used in full. Since these KOS don't have fully functioning subsets, sampling would render the data incomplete and hinder mapping to all relevant terms.

3.3.3 Evaluation of the Artifact

Evaluation is an integral part of DSR research, and artifacts must be evaluated with criteria that consider the context in which the artifact is implemented (Peffers et al., 2012). Sonnenberg and Brocke (2012) state that evaluations should occur throughout the design process to assess the artifact's progress as it is developed. They recommend selecting evaluation criteria based on the stage of the DSR process, which will inform the evaluation methods that can be used. In the case of an artifact, they propose four evaluation activities with appropriate criteria and methods. Table 3 outlines the activities, evaluation criteria, and evaluation methods that apply to creating the artifact in this research study.

Table 3

Evaluation Activities and Criteria

Activity	Input	Output (mandatory)	Eval. Criteria (exemplary)	Eval Methods (exemplary)

Eval. Activity 3	Instance of an artifact (prototype)	Validated artifact instance in an artificial setting (Proof of applicability)	Feasibility, ease of use, effectiveness, efficiency, fidelity with real-world phenomenon, operationality, robustness, suitability	Demonstration with the prototype, experiment with the prototype, experiment with the system, benchmarking, survey, expert interview, focus group.
Eval. Activity 4	Instance of an artifact	Validated artifact instance in a naturalistic setting (Proof of usefulness)	Applicability, effectiveness, efficiency, fidelity with real-world phenomenon, generality, impact on artifact environment and user, internal consistency, external consistency	Case study, field experiment, survey, expert interview, focus group

Note. Adapted from (Sonnenberg & vom Brocke, 2012)

Evaluation Activity 3 is completed to assess if an artifact works and how well it performs in what can be termed as prototyping or experimentation. Because the application context might be artificial, the proof may only demonstrate that the artifact applies to a task, in a system, or by a user. Evaluation Activity 4 demonstrates an artifact's usefulness and applicability in practice and should be embedded within an organization and tested with real tasks, systems, and users (Sonnenberg & vom Brocke, 2012). Peffers et al. (2012) also suggest ideal evaluation methods depending on the type of artifact. These methods (see Figure 9) align with those presented in Sonnenberg and vom Brocke's (2012) article.

Figure 9

Evaluation Methods based on Artifact

Logical Argument	An argument with face validity.
Expert Evaluation	Assessment of an artifact by one or more experts (e.g., Delphi study).
Technical Experiment	A performance evaluation of an algorithm implementation using real-world data, synthetic data, or no data, designed to evaluate the technical performance, rather than its performance in relation to the real world.
Subject-based Experiment	A test involving subjects to evaluate whether an assertion is true.
Action Research	Use of an artifact in a real-world situation as part of a research intervention, evaluating its effect on the real-world situation.
Prototype	Implementation of an artifact aimed at demonstrating the utility or suitability of the artifact.
Case Study	Application of an artifact to a real-world situation, evaluating its effect on the real-world situation.
Illustrative Scenario	Application of an artifact to a synthetic or real-world situation aimed at illustrating suitability or utility of the artifact.

Note. From Peffers et al., 2012

Based on these recommendations for evaluation, this research study most closely aligns with the evaluation's prototype and technical experiment methods. Therefore, after development, the utility or suitability of the artifact for the research questions outlined and evaluating its technical performance for the task of concept mapping will be demonstrated through showing results of implementation and results from experiments within the prototype and performance evaluations of the algorithms being tested. An overview of the research tasks and evaluation methods based on the recommendations above are shown in Table 4. The research study will not employ surveys, focus groups, or expert interviews at this stage.

Table 4

Task and Evaluation Activities

Task	**Evaluation**
Algorithm Implementation/ Semantic Analysis	Algorithm experiments on real-world covid vocabularies Algorithm performance evaluations
Evaluate Mapped Terms	Benchmarking - Comparison with the gold standard
Design Tool Prototype	Demo presentation of workflow in action
Annotation	Supervised learning approach – model accuracy validated with a test data set. Demonstrate annotation on clinical trials data

3.3.3.1 Algorithm implementation. Machine learning and natural language processing explore ways to automatically extract pertinent information from unstructured data. In particular, semantic analysis of the data using various models for entity recognition and classification tasks is a specific application of this field of artificial intelligence (Chowdhury, 2003). The mapping

problem can be addressed using classification models. These may range in complexity from simple similarity metrics and rule-based approaches, e.g., text search or regular expressions, to more complex ML models, e.g., support vector machines or naïve bayes. Development and implementation of the workflow involved a hybrid semantic analysis approach that combines similarity metrics with rule-based approaches and machine learning methods.

Precision and recall are standard evaluation metrics of the performance of classification models as evidenced in several of the solutions covered in the literature review (Allones et al., 2014; Kolyvakis et al., 2018; Zhou et al., 2012). Precision, recall, and F-measure calculated against the gold standard are used to compare the mappings and evaluate mapping accuracy. These model performance indicators are often used to assess and justify the use of certain machine learning models. They have an associated confusion matrix that provides visualization and description of their performance and whose values allow calculation of different metrics (Tharwat, 2020). The confusion matrix represents counts from predicted and actual values and appears as shown in Table 5.

Table 5

Confusion Matrix Example

		Predicted	
Actual		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

In this work, results are evaluated against a gold standard set of terms, therefore use of established measures such as those in Allones et al., (2014) and Zhou et al., (2012) which also was compared against a gold standard may be ideal. True positives are terms correctly

identified by the tool and present in the gold standard. False positives are terms incorrectly mapped. False negatives are terms not identified by the mapping but present in the gold standard, and true negatives are terms not identified by either the mapping or the gold standard.

The recall is the proportion of real positive cases that are correctly predicted positive (Powers, 2020). In this work as in the formula below and represents the percentage of mappings in the Gold Standard that were correctly identified:

$$Recall = \frac{\#\ correct\ found\ mappings}{\#\ all\ possible\ mappings}$$

Precision shows the proportion of predicted positives that are real positives (Powers, 2020) and represents the percentage of found mappings that agree with the gold standard. It is calculated as in the formula below:

$$Precision = \frac{\#\ correct\ found\ mappings}{\#\ all\ found\ mappings}$$

In addition to these measures, the F measure provides a weighted average of precision and recall. The closer the numbers are to 100%, the better the performance as measured by these metrics. This work uses the F-measure formula found in Allones (2014) to place more emphasis on precision rather than on recall where β=0.7, since that is more important in automated mapping task. The F-measure is calculated as

$$Fmeasure\ = 1 + \beta^2 \times \frac{precision\ \times\ recall}{(\beta^2\ \times precision)\ + recall}$$

Scores greater than 50% are ideal, with higher percentages indicating better performance. The confusion matrix can also be used to calculate the model's accuracy, which is a measurement of the proportion of correctly predicted terms out of all the terms.

3.3.3.2 Evaluation of Mapped Terms. A gold standard set of mappings generated by a team of domain experts is sometimes used to evaluate the result of a mapping solution (Allones et al., 2014; Fung et al., 2019; Gaudet-Blavignac et al., 2021; Hochheiser et al., 2016; Zhang & Bodenreider, 2007; Zhou et al., 2012). The study utilized benchmarking based on a gold standard set of mapping, rather than relying on domain experts at this stage. From the existing mappings within BioPortal, a validation set of terms were obtained for comparison with the mapping results. Mappings within BioPortal are generated both automatically and manually (Salvadores et al., 2013).

Some of the mappings have been generated through the NCBO's LOOM algorithm based on lexical matches between preferred names and a synonym. Others have been created through a UMLS unique concept identifier assigned by editors at the National Library of Medicine and others through OBO referencing. There are also mappings generated based on URI matches and those that are user-submitted. The author of the mapping will define the semantics of the mapping under consideration. Mapping relationships are usually identical, related, close or exact and are linked through the properties owl:sameAs, rdf:seeAlso, skos:relatedMatch, skos:closeMatch or skos:exactMatch (Salvadores et al., 2013).

For the evaluation, an API request to BioPortal returned 666 mappings between CIDO and the COVID-19 ontology, 489 mappings between COVID-19 ontology and LOINC, and 871 mappings between CIDO and LOINC. This dataset was then used as a gold standard for comparison with the results of the designed tool. The functioning of the tool was then assessed with a predetermined set of criteria. If a workflow segment produced half the number of similar matches, it would be considered partial functioning. An equal number of mappings would be considered similar or full functionality. A greater number of mappings would be considered to meet or exceed the gold standard.

Table 6

Criteria for Determining Functionality

Proportion of matches with the gold standard	Functionality Level
>50%	Partial
100%	Full
>100%	Met and exceeded

3.3.3.3 Annotation. Clinical trials have important information about diseases, drugs, labs, and other clinically relevant entities, which can be assessed through semantic analysis. A variety of natural language processing techniques including Named Entity Recognition (NER) can be used to locate and classify words into semantic categories. Machine learning-based NER features such as bag of words, sentence detection, part of speech tagging, dictionaries, etc., can be used to annotate text in addition to other rule-based or machine learning methods such as conditional random fields (CRF). These NER features were implemented in the workflow after which measures of performance for the model based on precision and recall were evaluated. These evaluation measures demonstrate an algorithm's practical use and performance (Junker et al., 1999).

3.3.3.4 Internal and External Validity. In design science research validity is addressed by assessing the solutions obtained. While statistical tests do have some utility for example, to assess the distribution of terms, overlap in terminology, or to evaluate coverage defined by distribution over match type (Bekhuis et al., 2013), these are not ideal for testing the validity of a designed artifact. According to Wieringa (2009), internal validity is determined by assessing whether the design implemented in a particular problem context satisfies the criteria identified in problem investigation through checking whether the solution has effects and if those effects

satisfy the criteria. Demonstrating that the novel workflow accomplishes the tasks outlined in the research questions and the results of the performance tests will determine whether the research has internal validity. Wieringa (2009) also recommends assessing whether the design implemented in a slightly different context satisfies the criteria to determine external validity. Thus, external validity is assessed using the workflow results for annotation of clinical trials. In addition, an additional clinical coding scheme, LOINC is tested in the workflow and its results evaluated for consistency.

3.4 Communicating Findings and Contributions to Knowledge

The final stage of the DSR process is to identify and analyze the theoretical contribution of the artifact and the design process (Lukka, 2003). Lukka (2003) notes that theoretical conclusions can be drawn regardless of whether the designed artifact was successful. Therefore, the results of the process should ideally be the development of a new theory or new knowledge that serves to refine an existing theory. Gregor and Hevner (2013) emphasize contributions to knowledge in the form of partial or incomplete theory. Alternatively, interesting, or empirical generalizations from the research can also be advanced.

Research Question 3 seeks to address theoretical conclusions by examining how the sociotechnical model can be leveraged to explain and assess mapping to achieve semantic interoperability. This research study focuses on mapping clinical coding schemes, which are crucial for the achievement of semantic interoperability. Sittig and Singh (2015) suggest that some sociotechnical models do not analyze and detail the technology element of their models in ways that can allow researchers to investigate the causes of HEALTH IT implementation and use problems or help identify specific solutions. Their eight-dimensional model (section 2.1.2) highlights clinical content as a dimension often overlooked with serious implications for health information systems. Clinical coding schemes provide a “cognitive interface between the inexact, subjective, highly variable world of biomedicine and the highly structured, tightly

controlled, digital world of computers" (Sittig & Singh, 2010). The challenges posed to clinicians by this aspect of technology are different than others. They can severely impact the clinician's workflow, patient satisfaction and safety, reduction of ambiguity in patient data, or development and implementation of decision support(Sittig et al., 2020). Understanding the issues and challenges involved with mapping can provide specific insights into the clinical content dimension of the technological component of these models.

Today, many applications and systems are using artificial intelligence applications, which sometimes replace or simulate humans' functions in certain ways. Being aware of the interdependencies between socio-technical dimensions is important for understanding how HEALTH IT is used in healthcare systems. Each dimension interacts and depends on one another and can positively or negatively impact another dimension of the system (Sittig & Singh, 2015). Therefore, the interplay between people, hardware and software, and clinical content dimensions may also be impacted by mapping and how it is achieved, especially as new information technologies are employed in the healthcare space. A final consideration is the impact of new and emerging diseases and their impact on the clinical content dimension and the external rules, regulations, and pressures dimension.

Some of these issues may be addressed by insights gained from the study. Challenges posed by these technological components can cause researchers to conclude wrongly that problems are due to hardware or software issues or that user error is at fault when more fine-grained issues related to implementing clinical vocabularies are at fault (Sittig & Singh, 2015). A discussion of the knowledge gained through the design process of this study and its specific impacts on certain dimensions of the sociotechnical model for health information technology is addressed in Chapter 5 in addition to some generalizations inferred from the research as it applies to the sociotechnical model.

Finally, this final stage of DSR methodology requires communication of the work and results (Kasanen et al., 1993; Peffers et al., 2008; Vaishnavi et al., 2015) Peffers (2008)

describes the final stage of the DSR as communication which involves sharing the problem and its import, the designed artifact, its utility, novelty, rigor, and effectiveness with researchers, relevant audiences, and practicing professionals. Presenting a completed document, a presentation, and a defense of the project to the dissertation committee meets this stated goal. Research paper publication, demonstrations and research talks can be given in future.

Chapter 4. Artifact Design and Evaluation

In previous sections the need for easily deployed and replicable methods for mapping was expressed. Reviews of the literature noted the difficulties and complexities involved with this process and the need for high level expertise and programming knowledge. Therefore, in this study one goal was to pursue the design of an artifact that could accomplish the mapping tasks with reasonable reliability as to be helpful especially in time-critical contexts where new clinical coding schemes are being developed to support lifesaving research, drug/vaccine development and various clinical applications.

In this section, a description of the datasets used to support the design of the novel workflow artifact will be provided first. Next the methods used to achieve the mapping outcomes will be described along with their results. Validation of those results against the gold standard set of terms will also be addressed. The study was performed by performing a series of operations over the data extracted from the clinical coding schemes, specifically those that were either created to deal with COVID-19 or coronavirus infectious diseases more generally, these include the COVOC Coronavirus Vocabulary, the COVID-19 Ontology and the Coronavirus Infectious Disease Ontology.

4.1 Description of the study data

The clinical coding schemes used in this study vary in their content and use a variety of properties to express common relations and attributes. For example, to express important hierarchical relationships among class and subclass terms and to other vocabularies, the ontologies used properties such as:

```
rdfs:subClassOf
skos:narrower,
is_a
oboInOwl:hasDBXref,
```

skos:ExactMatch

skos:closeMatch

The properties that represent class, subclass and property term labels can vary, for example the most common is rdfs:label, however, each vocabulary may use their own preferred label form or use the skos:prefLabel property to identify the preferred term. The clinical schemes also include properties to store synonyms such as oboInOwl:hasRelatedSynonym are also multiple and varied and don't follow the recommendation of a single standard such as SKOS.

The namespaces included in the schemes can give an idea of the commonly reused properties which gives an indication of interoperability, and which help to align these ontologies with the FAIR Guiding Principles proposition that all research data should be Findable, Accessible, Interoperable and Reusable (FAIR) for both machine and human users (Wilkinson et al., 2016). Figure 10 provides a list of namespaces across the vocabularies along with an indication of their commonality highlighted. The most common namespaces of note include dc, skos, rdfs, which indicate an intention to align with specific schema recommendations and chebi, obo, owl, mondo, and ncbitaxon which give an indication of the domain of these schemes.

To perform the mappings, the focus was placed mostly on the ontology terms identified by the unique rdfs:label that accompanies them in the ontology, as well as definitions of each scheme. Table 7 provides a summary of the breakdown of these values from each vocabulary. The Coronavirus infectious disease ontology had the most terms, properties, and definitions, it uses coronavirus terms from existing reliable reference ontologies that align with OBO Foundry principles, and under the Basic Formal Ontology (BFO), an ISO/IEC standard 21838-2 (https://www.iso.org/standard/74572.html) top-level ontology (He et al., 2020) which makes it highly interoperable and a good fit for enhancing other newly developed schemes through mapping.

Figure 10

Common Namespaces across Datasets

CIDO	COVOC	COVID-19
dc:"http://purl.org/dc/elements/1.1/"	dc:"http://purl.org/dc/elements/1.1/"	xml:base:"https://bio.scai.fraunhofer.de/ontology/COVID.owl"
nbo:"http://purl.obolibrary.org/obo/nbo.owl#"	go:"http://purl.obolibrary.org/obo/go#"	dc:"http://purl.org/dc/elements/1.1/"
obo:"http://purl.obolibrary.org/obo/"	hp:"http://purl.obolibrary.org/obo/hp#"	efo:"http://www.ebi.ac.uk/efo/"
owl:"http://www.w3.org/2002/07/owl#"	efo:"http://www.ebi.ac.uk/efo/"	obo:"http://purl.obolibrary.org/obo/"
rdf:"http://www.w3.org/1999/02/22-rdf-syntax-ns#"	obo:"http://purl.obolibrary.org/obo/"	owl:"http://www.w3.org/2002/07/owl#"
www:"http://www.referent-tracking.com/"	owl:"http://www.w3.org/2002/07/owl#"	rdf:"http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xml:"http://www.w3.org/XML/1998/namespace"	rdf:"http://www.w3.org/1999/02/22-rdf-syntax-ns#"	xml:"http://www.w3.org/XML/1998/namespace"
xsd:"http://www.w3.org/2001/XMLSchema#"	xml:"http://www.w3.org/XML/1998/namespace"	xsd:"http://www.w3.org/2001/XMLSchema#"
core:"http://purl.obolibrary.org/obo/core#"	xsd:"http://www.w3.org/2001/XMLSchema#"	foaf:"http://xmlns.com/foaf/0.1/"
doap:"http://usefulinc.com/ns/doap#"	core:"http://purl.obolibrary.org/obo/uberon/core#"	rdfs:"http://www.w3.org/2000/01/rdf-schema#"
foaf:"http://xmlns.com/foaf/0.1/"	foaf:"http://xmlns.com/foaf/0.1/"	skos:"http://www.w3.org/2004/02/skos/core#"
rdfs:"http://www.w3.org/2000/01/rdf-schema#"	ncit:"http://purl.obolibrary.org/obo/ncit#"	chebi:"http://purl.obolibrary.org/obo/chebi/"
skos:"http://www.w3.org/2004/02/skos/core#"	pato:"http://purl.obolibrary.org/obo/pato#"	covid:"https://bio.scai.fraunhofer.de/ontology/covid#"
chebi:"http://purl.obolibrary.org/obo/chebi/"	rdfs:"http://www.w3.org/2000/01/rdf-schema#"	mondo:"http://purl.obolibrary.org/obo/mondo#"
terms:"http://purl.org/dc/terms/"	skos:"http://www.w3.org/2004/02/skos/core#"	terms:"http://purl.org/dc/terms/"
NDF-RT:"http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#"	chebi:"http://purl.obolibrary.org/obo/chebi/"	protege:"http://protege.stanford.edu/plugins/owl/protege#"
uberon:"http://purl.obolibrary.org/obo/uberon#"	core1:"http://purl.obolibrary.org/obo/core#"	oboInOWL:"http://www.geneontology.org/formats/oboInOWL#"
protege:"http://protege.stanford.edu/plugins/owl/protege#"	covoc:"http://purl.obolibrary.org/obo/covoc/"	apollo_sv:"http://purl.obolibrary.org/obo/apollo_sv.owl/"
taxslim:"http://purl.obolibrary.org/obo/ncbitaxon/subsets/taxslim#"	mondo:"http://purl.obolibrary.org/obo/mondo#"	ncbitaxon:"http://purl.obolibrary.org/obo/ncbitaxon#"
oboInOwl:"http://www.geneontology.org/formats/oboInOwl#"	terms:"http://purl.org/dc/terms/"	oboInOwl1:"http://ontofox.hegroup.org/vkdjgd5o.owl#oboInOwl:"
apollo_sv:"http://purl.obolibrary.org/obo/apollo_sv.owl/"	chebi2:"http://purl.obolibrary.org/obo/chebi#2"	oboInOwl2:"https://bio.scai.fraunhofer.de/ontology/covid#oboInOwl:"
ncbitaxon:"http://purl.obolibrary.org/obo/ncbitaxon#"	chebi3:"http://purl.obolibrary.org/obo/chebi#3"	owl:Ontology rdf:about:"https://bio.scai.fraunhofer.de/ontology/COVID.owl"
	chebi4:"http://purl.obolibrary.org/obo/chebi#1"	
	ubprop:"http://purl.obolibrary.org/obo/ubprop#"	
	cellline:"http://www.ebi.ac.uk/cellline/"	
	oboInOwl:"http://www.geneontology.org/formats/oboInOwl#"	
	patterns:"http://www.co-ode.org/patterns#"	
	ncbitaxon:"http://purl.obolibrary.org/obo/ncbitaxon#"	

Table 7

Metrics for Clinical Coding Schemes Used

Vocabulary	# of ontology terms	# with definition	# of properties
COVID-19 Vocabulary (COVOC)	547	373	41
Coronavirus Infectious Disease Ontology (CIDO)	7866	3998	451
COVID19 Ontology	2278	1400	8

Reuse of ontologies is a critical aspect of their existence as a means of knowledge representation, however searching for and identifying concepts and predicates is a tedious and time-consuming process (Katsumi et al., 2016). To aid in this process the rdf/xml or rdf/ttl formats of these ontologies were obtained and loaded into KNIME with the Triple File Reader node and access to the contents made available for querying with the SPARQL Insert node which added the triples to an in-memory semantic web endpoint. SPARQL queries for viewing all classes and associated labels, definitions and properties were used to display the results shown in the table. Here is an example which demonstrates loading the triple data into the memory endpoint within the workflow and querying non-duplicate terms from it.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: http://www.w3.org/2004/02/skos/core#

SELECT DISTINCT ?s ?label
```

```
WHERE
{
?s a owl:Class .
?s rdfs:label ?label .
}
```

The query output was passed to the nodes and processes for mapping within the workflow. The data used for validation was accessed through the BioPortal REST API and was downloaded in JSON format, it consists of mappings between the CIDO and COVID-19 ontologies, COVID-19 and LOINC, and mappings between CIDO and LOINC (see Table 10).

4.2 Artifact Design and Implementation – DSR

The major design of the workflow was aimed at answering research question 1 which addresses the utility and functioning of the mapping task with a node-based workflow process. This stage reflected the views of Vaishnavi et al. (2015) about the iterative nature of a process where the problems are imperfectly understood and multifaceted. Therefore, various solutions were explored and reassessed where not found to be appropriate. The mapping results were obtained through three testing processes. These have been identified and named as the Lexical Series Matcher, Document Classification Matcher, and Semantic Meaning Matcher implemented via workflow nodes in the design tool. The flowchart in Figure 11 gives an overview of the mapping workflow segments.

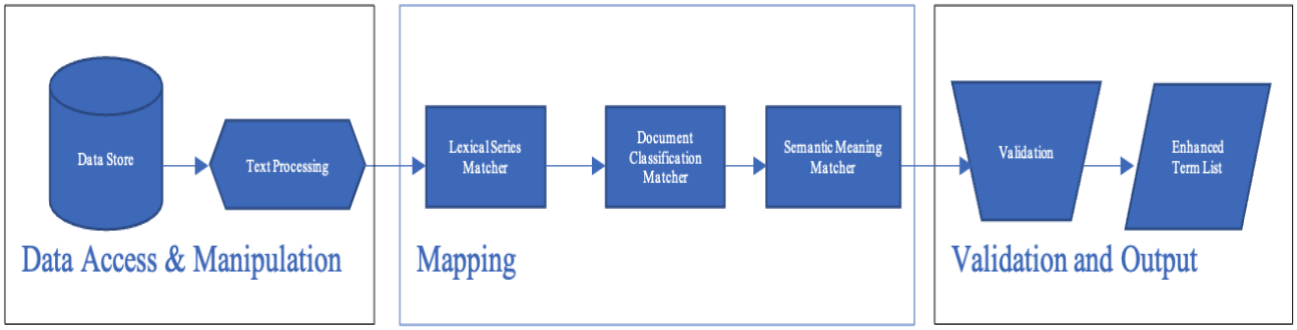
4.3 Design Tool

The research artifact was created using the KNIME Analytics Platform which is a free and open-source platform for data analytics, reporting and the integration of various components for machine learning and data mining through modular data pipelining via a graphical user interface (GUI). In the GUI nodes are assembled to combine different data

sources. The nodes cover tasks such as preprocessing, data analysis and visualization without programming or with minimal programming involved.

Figure 11

Flowchart Illustrating Workflow Artifact Segments



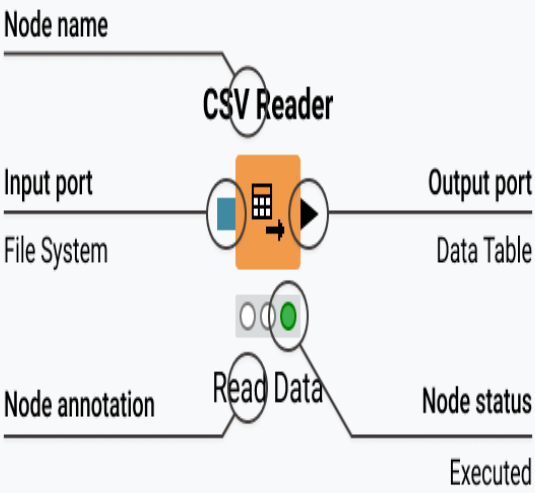
4.3.1 KNIME Nodes

Each task within the tool is represented by a node. A node is shown as a colored box with input and output ports, representing the data the node will process and the resulting data after processing (See Figure 12). They can perform a variety of tasks including reading/writing files, transforming data, training models, creating visualizations and more. However, each node has specific settings which must be adjusted in a configuration dialog to align with their functioning and the task's specific objective.

Nodes have four states including, not configured, configured, executed and error, represented by a traffic light. A series of connected nodes defines a workflow, and they are connected to each other via their input and output ports, once the workflow is executed, data inside the workflow flows from left to right through the connections. Nodes can also be combined into components – nodes that contain a sub-workflow; and metanodes - which allows collapsing and organizing the workflow into sections that make it easier for people to understand the structure a bit more.

Figure 12

KNIME Nodes and Configuration Example



This work utilizes nodes that enable data blending, transformation, machine learning, and visualization. Three mapping scenarios were implemented for the datasets, in the first, the lexical series matcher uses joiner nodes in combination with nodes which calculate string distance and string similarity to identify potential term mappings based on the URI value matches or term similarity across table. The document classification – term definition matcher implements a model which computes cosine similarity through the document similarity learner and document similarity predictor nodes. Term definitions are transformed to produce document vectors that inform the model thus predict which definitions are closely matching. Finally, the Semantic Similarity matcher uses nodes for data input and transformation and a python script for testing the semantic similarity of the terms with the python Scispacy library.

4.4 Artifact Development – Mapping Workflows

Vaishnavi and Keuchler (2015) recommend various suggestion and development patterns for design science research when determining the strategies that can be employed to develop a solution to the research problem and in generating knowledge that is of general value. The *Preliminaries Type* pattern where tools and techniques useful for the problem space are identified and used guides the answers to the research questions. This is recommended when the motivation is to expand the choice of tools and techniques that can be used as the solution to the research problem.

Although use of the traditional tools common to the research space is one possible pattern, approaching the issue from the problem space perspective, allows the researcher to use their knowledge of tools and techniques to see whether a promising method has been overlooked by the research community (Vaishnavi et al., 2015). From the pilot it was determined that testing ETL tools for mapping and the possibility that these tools do in fact support this process was worth pursuing. The intent is to demonstrate that the ETL tool can support mapping tasks and can also offer some efficiencies and opportunities not available with current solutions. Therefore, in this section an explanation of the various matchers deployed within the workflow tool will be addressed.

Figure 12 shows an example of a data input node. In this work data input is accomplished with the Triple File Reader from which a SPARQL insert passes the ontology triples to an in-memory endpoint for SPARQL querying. Extracted data is passed to the Lexical Series Matcher, Document Annotation – Term Definition Matcher, and Semantic Similarity Matcher for processing. Each workflow segment was created and refined in several iterations. Outputs include the workflow artifact itself, and an enriched table of mappings from the high performing mapping segments which is then used in the clinical trial annotation workflow segment. The mapping workflow segments are further described in the following sections.

4.4.1 Lexical Series Matcher

In this matcher, two types of matching processes occurred. Here follows a description of the mapping results between COVOC and CIDO. The two datasets were cleaned and normalized with various text processing nodes. To find URI matches, a joiner node was added to the workflow and configured to find terms with shared URIs. Terms with the same URI are an indication that the concept in question is identical. Matches with the same URI were given the skos:exactMatch as match type label. A skos:exactMatch is a type of link which indicates a high degree of confidence that two concepts can be used interchangeably (Miles & Bechhofer, 2009). Between COVOC and CIDO there were (n=98) terms that were assigned skos:exactMatch after removing duplicates. Between CIDO and the COVID-19 ontology there were (n=586) terms with the same URI assigned to skos:exactMatch after removing duplicates.

The next step in the lexical series matcher involved the employment of string similarity algorithms. In KNIME, this is set up through the string distance node, similarity learner and similarity search node. In the string distance node, we configure the algorithmic settings for each algorithm. The algorithms used to select and configure the distances used for measurement included the Jaro-Winkler Distance which is common distance measure for the difference between two strings. The distances range from 0 to 1 where 0 means the strings are equal and 1 means no similarity between the strings. For example, with the term inputs:

```
Input: t1 = “myelopathy”, t2 = “lymphocyte”
Output: Jaro Similarity = 0.27917

Input: t1 = “expectorant”, t2 = “expectorate”
Output: Jaro Similarity = 0.03636
```

The N-gram Tversky distance which provides a probabilistic model for relations between neighbored letters by predicting the next item in a sequence of items. This algorithm computes the number of n-grams from each character or word in two strings. The distance is computed by dividing the number of similar n-grams by the maximal number of n-grams. Example:

```
Input: t1 = “pulmonary embolism”, t2 = “pulmonary edema”
Output: N-gram Tversky = 0.10667

Input: t1 = “regulation of actin cytoskeleton organization”, t2= “regulation
of actin cytoskeleton reorganization”
Output N-gram Tversky = 0.01296
```

Finally, the Levenshtein distance which counts the minimum number of edit operations needed to transform one string into another, where an operation is defined as an insertion, deletion or substitution of a single character or a transposition of two adjacent characters. Example:

```
Input: t1 = “specimen from organism”, t2 = “specific granule”
Output: Levenshtein = 0.54545

Input: t1 = “recurrent lower respiratory tract infections”, t2 = “recurrent
upper respiratory tract infections”
Output: Levenshtein = 0.06818
```

In the lexical matcher, the results of these algorithms are combined into a single set of matches with duplicates removed. The match types skos:closeMatch which indicates that two concepts are sufficiently similar that they can be used interchangeably was applied to lexical matches with string distances equal to 0. For terms with a string distance between 0.01 and 0.25, it was noted they can either be similar or dissimilar, however differences were due to things like variant spellings or compound terms. This was especially true with LOINC terms as in Figure 15.

Figure 13

Mapping Output between the clinical coding schemes COVOC and CIDO

COVOC_URI	COVOC TERM	CIDO_URI	CIDO TERM
http://purl.obolibrary.org/obo/HP_0012378	fatigue	http://purl.obolibrary.org/obo/HP_0012378	fatigue
http://purl.obolibrary.org/obo/NCBITaxon_31631	human coronavirus oc43	http://purl.obolibrary.org/obo/NCBITaxon_31631	human coronavirus oc43
http://purl.obolibrary.org/obo/PATO_0000169	viability	http://purl.obolibrary.org/obo/PATO_0000169	viability
http://purl.obolibrary.org/obo/HP_0002018	nausea	http://purl.obolibrary.org/obo/HP_0002018	nausea
http://purl.obolibrary.org/obo/NCBITaxon_9838	camelus dromedarius	http://purl.obolibrary.org/obo/NCBITaxon_9838	camelus dromedarius
http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
http://purl.obolibrary.org/obo/CHEBI_44032	indinavir	http://purl.obolibrary.org/obo/CHEBI_44032	indinavir
http://purl.obolibrary.org/obo/HP_0031246	nonproductive cough	http://purl.obolibrary.org/obo/HP_0031246	nonproductive cough
http://purl.obolibrary.org/obo/NCBITaxon_694009	severe acute respiratory syndrome-related coronavirus	http://purl.obolibrary.org/obo/NCBITaxon_694009	severe acute respiratory syndrome-related coronavirus
http://purl.obolibrary.org/obo/NCBITaxon_694009	severe acute respiratory syndrome-related coronavirus	http://purl.obolibrary.org/obo/NCBITaxon_694009	sars-cov

Figure 14

Mapping Output between Clinical Coding Schemes CIDO and COVID-19

S CIDO-URI	S TERM	D distance	S COV19-URI	S TERM (right)
http://purl.obolibrary.org/obo/HP_0004430	severe combined immunodeficiency	0	http://purl.obolibrary.org/obo/DOID_627	severe combined immunodeficiency
http://purl.obolibrary.org/obo/CIDO_0000365	personal protective equipment	0	http://purl.obolibrary.org/obo/OMIT_0001154	personal protective equipment
http://purl.obolibrary.org/obo/HP_0001649	tachycardia	0	http://purl.obolibrary.org/obo/SYMP_0000529	tachycardia
http://purl.obolibrary.org/obo/HP_0002014	diarrhea	0	http://purl.obolibrary.org/obo/SYMP_0000570	diarrhea
http://purl.obolibrary.org/obo/HP_0012819	myocarditis	0	http://purl.obolibrary.org/obo/SYMP_0000095	myocarditis
http://purl.obolibrary.org/obo/HP_0012531	pain	0	http://purl.obolibrary.org/obo/SYMP_0000099	pain
http://purl.obolibrary.org/obo/HP_0000819	diabetes mellitus	0	http://purl.obolibrary.org/obo/DOID_9351	diabetes mellitus
http://purl.obolibrary.org/obo/HP_0001254	lethargy	0	http://purl.obolibrary.org/obo/SYMP_0000075	lethargy
http://purl.obolibrary.org/obo/HP_0100749	chest pain	0	http://purl.obolibrary.org/obo/SYMP_0000576	chest pain
http://purl.obolibrary.org/obo/HP_0004756	ventricular tachycardia	0	http://purl.obolibrary.org/obo/SYMP_0000827	ventricular tachycardia
http://purl.obolibrary.org/obo/UBERON_0007311	sputum	0	http://purl.obolibrary.org/obo/SYMP_0000431	sputum
http://purl.obolibrary.org/obo/HP_0001663	ventricular fibrillation	0	http://purl.obolibrary.org/obo/SYMP_0000899	ventricular fibrillation
http://purl.obolibrary.org/obo/OBI_0002608	oropharyngeal swab specimen	0	http://purl.obolibrary.org/obo/NCIT_C155835	oropharyngeal swab specimen
http://purl.obolibrary.org/obo/HP_0012337	abnormal homeostasis	0	http://purl.obolibrary.org/obo/OGMS_0000037	abnormal homeostasis
http://purl.obolibrary.org/obo/HP_0012378	fatigue	0	http://purl.obolibrary.org/obo/SYMP_0019177	fatigue
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N...	methylprednisolone	0	http://purl.obolibrary.org/obo/NCIT_C647	methylprednisolone
http://purl.obolibrary.org/obo/PR_000036009	angiotensin ii	0	http://purl.obolibrary.org/obo/CHEBI_48432	angiotensin ii
http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT.owl#N...	chloroquine phosphate	0	http://purl.obolibrary.org/obo/NCIT_C47445	chloroquine phosphate
http://purl.obolibrary.org/obo/HP_0000822	hypertension	0	http://purl.obolibrary.org/obo/DOID_10763	hypertension
http://purl.obolibrary.org/obo/HP_0001888	lymphopenia	0	http://purl.obolibrary.org/obo/DOID_614	lymphopenia
http://purl.obolibrary.org/obo/CIDO_0000280	viral life cycle	0	http://purl.obolibrary.org/obo/GO_0019058	viral life cycle
http://purl.obolibrary.org/obo/HP_0011675	arrhythmia	0	http://purl.obolibrary.org/obo/SYMP_0000287	arrhythmia
http://purl.obolibrary.org/obo/CIDO_0000368	n95 respirator	0	https://bio.scai.fraunhofer.de/ontology/COVID_0000139	n95 respirator
http://purl.obolibrary.org/obo/HP_0002315	headache	0	http://purl.obolibrary.org/obo/SYMP_0000504	headache
http://purl.obolibrary.org/obo/HP_0001627	abnormal heart morphology	0	http://purl.obolibrary.org/obo/MP_0000266	abnormal heart morphology
http://purl.obolibrary.org/obo/HP_0001250	seizure	0	http://purl.obolibrary.org/obo/SYMP_0000124	seizure
http://purl.obolibrary.org/obo/HP_0012622	chronic kidney disease	0	http://purl.obolibrary.org/obo/DOID_784	chronic kidney disease
http://purl.obolibrary.org/obo/HP_0025143	chills	0	http://purl.obolibrary.org/obo/SYMP_0019174	chills

Figure 15

Mapping Output between Clinical Coding Schemes CIDO and LOINC

S CIDO-URI	S TERM	D distance	S LOINC-URI	S TERM (right)
http://purl.obolibrary.org/obo/CHEBI_23981	ethanolamines	0.043	http://purl.bioontology.org/ontology/LNC/LP15562-9	ethanolamine
http://purl.obolibrary.org/obo/CHEBI_17359	sulfite	0.077	http://purl.bioontology.org/ontology/LNC/MTHU035...	sulfites
http://purl.obolibrary.org/obo/SO_0001537	structural_variant	0.118	http://purl.bioontology.org/ontology/LNC/LA26802-1	structural variant
http://purl.obolibrary.org/obo/UBERON_00...	coronary artery	0.111	http://purl.bioontology.org/ontology/LNC/LP34720-0	coronary arteries
http://purl.obolibrary.org/obo/CHEBI_28262	dimethyl sulfoxide	0.091	http://purl.bioontology.org/ontology/LNC/LP18068-4	dimethylsulfoxide
http://purl.obolibrary.org/obo/PR_P20309	muscarinic acetylcholine receptor m3 (human)	0.105	http://purl.bioontology.org/ontology/LNC/LP24772...	muscarinic acetylcholine receptor m3 ab
http://purl.obolibrary.org/obo/GO_0005245	voltage-gated calcium channel activity	0.121	http://purl.bioontology.org/ontology/LNC/LP40151-0	voltage-gated calcium channel ab
http://purl.obolibrary.org/obo/NCBITaxon_...	porcine epidemic diarrhea virus cv777	0.077	http://purl.bioontology.org/ontology/LNC/MTHU056...	porcine epidemic diarrhea virus
http://purl.obolibrary.org/obo/NCBITaxon_...	porcine epidemic diarrhea virus br1/87	0.104	http://purl.bioontology.org/ontology/LNC/MTHU056...	porcine epidemic diarrhea virus
http://purl.obolibrary.org/obo/CHEBI_22718	benzoates	0.067	http://purl.bioontology.org/ontology/LNC/MTHU008...	benzoate
http://purl.obolibrary.org/obo/CHEBI_24828	indoles	0.091	http://purl.bioontology.org/ontology/LNC/MTHU012...	indole
http://purl.obolibrary.org/obo/NCBITaxon_...	porcine deltacoronavirus sichuan	0.154	http://purl.bioontology.org/ontology/LNC/MTHU056...	porcine deltacoronavirus
http://purl.obolibrary.org/obo/PR_000013...	trypsin-1	0.143	http://purl.bioontology.org/ontology/LNC/MTHU012...	trypsin
http://purl.obolibrary.org/obo/UBERON_00...	limb	0.143	http://purl.bioontology.org/ontology/LNC/LP7395-9	limbs

Note. The matches show variant spellings in the results.

Figure 16

Mapping Output between Clinical Coding Schemes COVOC and CIDO

COVOC_URI	COVOC TERM	distance	CIDO_URI	CIDO TERM
http://purl.obolibrary.org/obo/MONDO_0005618	anxiety disorder	0	http://purl.obolibrary.org/obo/DOID_2030	anxiety disorder
http://purl.obolibrary.org/obo/COVOC_0030021	methylprednisolone	0	http://purl.obolibrary.org/obo/CHEBI_6888	methylprednisolone
http://purl.obolibrary.org/obo/COVOC_0030021	methylprednisolone	0	http://purl.obolibrary.org/obo/CHEBI_6888	6alpha-methylprednisolone
http://purl.obolibrary.org/obo/NCIT_C25201	sensitivity	0	http://purl.obolibrary.org/obo/OBCS_0000058	sensitivity
http://www.ebi.ac.uk/efo/EFO_0000684	respiratory system disease	0	http://purl.obolibrary.org/obo/DOID_1579	respiratory system disease
http://purl.obolibrary.org/obo/MONDO_0005108	viral infectious disease	0	http://purl.obolibrary.org/obo/DOID_934	viral infectious disease
http://purl.obolibrary.org/obo/NCIT_C17021	protein	0	http://purl.obolibrary.org/obo/PRO_000000001	protein
http://purl.obolibrary.org/obo/CHEBI_36080	protein	0	http://purl.obolibrary.org/obo/PRO_000000001	protein
http://www.ebi.ac.uk/efo/EFO_0004420	genome	0	http://purl.obolibrary.org/obo/OGG_0000000001	genome
http://purl.obolibrary.org/obo/NCIT_C155831	nasopharyngeal swab specimen	0	http://purl.obolibrary.org/obo/OBI_0002606	nasopharyngeal swab specimen

Note. These results can be assigned the semantic label skos:closeMatch

Smaller distances usually indicated a higher likelihood of the term being the same or similar, however this was not a rule. Some terms with small edit distances, particularly in the results between COVOC and CIDO were not similar beyond using the same alphanumeric characters.

Therefore, matches within and above this range must be reviewed before they can be accepted. Terms with string distances above 0.25 were not considered. Figure 19 shows a sample of mappings between terms with a string distance of 0. For example, the terms methylprednisolone and 6-alpha-methylprednisolone are identified as matched with a distance score of zero even though the characters have some difference, however in the source ontology CHEBI both terms have same URI which indicates that the match is correct. A summary of the output of the lexical series matcher for terms with string distances less than 0.25 across vocabularies is provided in Table 8.

Table 8

Results of Lexical Series Matcher

Match Type	COVOC/ CIDO	COVOC/ COVID19	COVOC/ LOINC	CIDO/ COVID19	CIDO/ LOINC	COVID19 / LOINC
skos:exactMatch	98	94	0	586	0	0
skos:closeMatch	48	53	121	28	424	245
Match (Review)	346	38	23	289	323	153
Total Mappings <0.25	492	185	144	903	747	398

4.4.2 Document Annotation – Term Definition Matcher

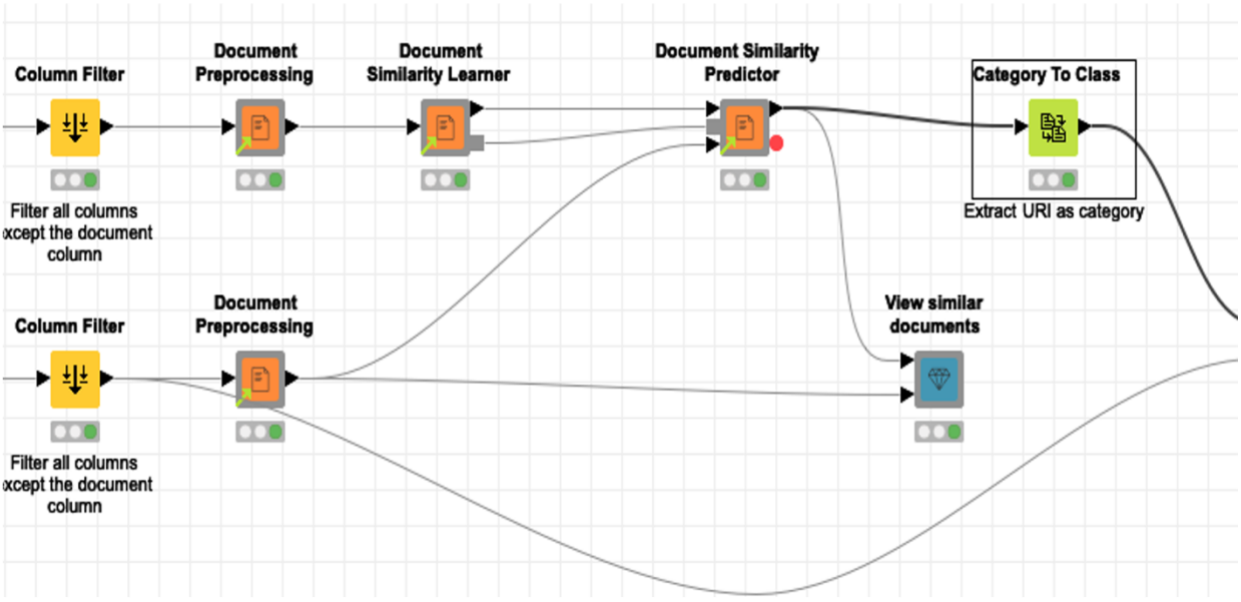
String similarity algorithms perform their operation on the entity name labels only and similarly named entities in the other dataset. This technique therefore analyzes the terms

independent of their assigned definitions and relationships with other terms. In addition, a word can be a homonym, where it's meaning is different depending on the usage context.

4.4.2.1 Matching process. For this next phase of the matching process, the intent is to assess the definitions of the terms to identify matches with the document similarity learner component which takes as takes input a corpus of documents via a preprocessing component and provides as output a model to be used with the document similarity predictor component.

Figure 17

Subsection of Document Similarity Predictor



4.4.2.2 Similarity predictor: Cosine similarity, from 1 to 0. The document similarity learner component uses nodes that creates a document vector for each document or definition in this case, representing it in the terms space to create a bag of words. The document similarity predictor then applies the model obtained by the Document similarity learner to a test document, in this case the target definitions. It computes the cosine similarity between the original corpus of definitions table and the test definitions table.

Cosine similarity is a metric that quantifies the similarity between two or more vectors. Vectors are typically non-zero and within an inner product space. Cosine similarity measures

the angle between the two vectors projected in a multi-dimensional space. As the measurement gets closer to 1 then the angle between vectors is smaller. Therefore, these components will convert our definitions into words or phrases within a document or sentence into a vectorized form of representation which is then used within the cosine similarity formula to obtain a measurement of similarity. If the cosine similarity is 1 it implies that the two definitions are exactly alike. If the cosine similarity is 0 between definitions means, there are no similarities.

4.4.2.3 Summarized Thresholds. Not all class terms in a coding scheme include term definitions, however, those definitions that existed were extracted and included in the matcher. The number of annotations available in each coding scheme is available beside the scheme name in the table in superscript. As expressed above similarity scores range from 1 to 0 with 1 predicted to be a skos:exactMatch. On review, it was determined that definitions with a similarity score between 0.95 and 0.99 could be considered *close matches.* Term definitions with similarity scores between 0.93 and 0.7 are sometimes related but sometimes not, they require expert review to determine the type. See for example in Figure 19 the term definitions with a similarity score of 0.91. Similarity scores below 0.6 are unlikely to be related in any way to each other and were not considered.

Figure 18

Mapping Output from Document Similarity Matcher

CIDO_URI	Automatically Preprocessed Document	COV19_URI	Matching COVID-19 definition	similarity
http://purl.obolibrary.org/obo/UBERON_0011216	"subdivision anatomical system"	http://purl.obolibrary.org/obo/UBERON_0011216	"subdivision anatomical system"	1
http://purl.obolibrary.org/obo/CHEBI_15939	"triterpenoid saponin glucosiduronide derivative 3beta-hydroxy-11-oxoolean-12-en-30-oic acid"	http://purl.obolibrary.org/obo/CHEBI_15939	"triterpenoid saponin glucosiduronide derivative 3beta-hydroxy-11-oxoolean-12-en-30-oic acid"	1
http://purl.obolibrary.org/obo/UBERON_0004111	"tubepassage connects distinct anatomical spaces"	http://purl.obolibrary.org/obo/UBERON_0004111	"tubepassage connects distinct anatomical spaces"	1
http://purl.obolibrary.org/obo/UBERON_0001981	"vessel blood circulates body"	http://purl.obolibrary.org/obo/UBERON_0001981	"vessel blood circulates body"	1
http://purl.obolibrary.org/obo/UBERON_0004923	"wall organ forms layer"	http://purl.obolibrary.org/obo/UBERON_0004923	"wall organ forms layer"	1
http://purl.obolibrary.org/obo/HP_0002086	"abnormality respiratory systeminclude airwayslungsrespiratory muscles"	http://purl.obolibrary.org/obo/HP_0002086	"abnormality respiratory systeminclude airwayslungsrespiratory muscles"	0.9999999999999999
http://purl.obolibrary.org/obo/UBERON_0000062	"anatomical structure performs specific function functions wp"	http://purl.obolibrary.org/obo/UBERON_0000062	"anatomical structure performs specific function functions wp"	0.9999999999999999
http://purl.obolibrary.org/obo/CHEBI_55438	"angiotensin compound consisting	http://purl.obolibrary.org/obo/CHEBI_55438	"angiotensin compound consisting	0.9999999999999999

Figure 19

Example of Related Definitions in the same Subclass

The output of the document similarity workflow across all vocabularies for various term definitions by similarity measure is presented in Table 9 below.

Table 9

Number of Mapped Term Types for Document Similarity Matcher

Similarity Score **Coding Scheme**	**1** **skos:exactMatch**	**>=0.94 & <=0.99** **skos:closeMatch**	**>=0.7 & <=0.93** **Check Match**
COVOC$^{(n=373)}$| CIDO$^{(n=1400)}$	18	119	383
CIDO$^{(n=3998)}$ | COVID19$^{1(n=1400)}$	79	8	502

COVOC $^{(n=373)}$| COVID19$^{(n=1400)}$	18	7	59

Note. # of definitions in scheme added in superscript

4.4.3 Semantic Similarity Matcher

Another way to consider the relatedness between terms is by considering the semantic meaning of terms. The primary task of this workflow segment is to compute the semantic similarities of clinical coding scheme terms with each other. We experiment with a measure of semantic relatedness using soft cosine for prediction of possible relatedness.

4.4.3.1 Soft Cosine Measure. The soft cosine measure is a machine learning method that allows for assessing the similarity between two documents, even when there are no words in common. It uses a measure of similarity between words which are obtained through word2vec vector embeddings of words and has been demonstrated to outperform many semantic text similarity tasks. By modeling synonymy, even when sentences have no words in common, the soft cosine measure can calculate the similarity between sentences (Sidorov et al., 2014).

4.4.3.2 Implementation. The term labels used for this workflow segment were extracted from the triple file and passed to the python script node. This node allows executing a python script in a local python environment. Input and output ports can be dynamically added as needed for passing data into and out of the executable script in this case the soft cosine algorithm. The output is then parsed and aligned with the original datasets to be able to view and filter the predicted matches. To improve the performance of the model, instead of using a set of general terns, a spacy pipeline for biomedical data was used in the algorithm.

4.4.3.3 Semantic Matcher Output. Mapping results from the semantic matcher were filtered to only show and assess the results for terms with a similarity score above 0.95. Additionally, the suggested mappings with the same URI from this workflow were also isolated

as a means of checking the range of similarity scores within the subset and checking whether that gave a better ability to guess what similarity scores were likely to produce reliable mappings. The semantic matcher returns a score for every term against every other term in the set, therefore it has a high computing cost specifically for any clinical coding scheme with thousands of terms to be calculated. While the term labels in all cases appear to be the same, the definitions assigned to the terms can have more or less detail provided as in Figure 20.

Figure 20

High Similarity Scored Terms from Semantic Matcher

amino acid	A carboxylic acid containing one or more amino groups.	http://purl.obolibrary.org/obo/CHEBI_33709	amino acid	Any amino acid whose side chain is capable of forming one or more hydrogen bonds.	http://purl.obolibrary.org/obo/CHEBI_26167	1
amino acid	A carboxylic acid containing one or more amino groups.	http://purl.obolibrary.org/obo/CHEBI_33709	amino acid	Any monocarboxylic acid which also contains a separate (alcoholic or phenolic) hydroxy substituent.	http://purl.obolibrary.org/obo/CHEBI_35868	1
amino acid	A carboxylic acid containing one or more amino groups.	http://purl.obolibrary.org/obo/CHEBI_33709	amino acid	An oxoacid containing a single carboxy group.	http://purl.obolibrary.org/obo/CHEBI_25384	1
amino acid	A carboxylic acid containing one or more amino groups.	http://purl.obolibrary.org/obo/CHEBI_33709	amino acid	Any aromatic carboxylic acid that consists of benzene in which at least a single hydrogen has been substituted by a carboxy group.	http://purl.obolibrary.org/obo/CHEBI_22723	1
chest pain	An unpleasant sensation characterized by physical discomfort (such as pricking, throbbing, or aching) localized to the chest.	http://purl.obolibrary.org/obo/HP_0100749	Chest pain	An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.	http://purl.obolibrary.org/obo/HP_0012531	1
abdominal pain	An unpleasant sensation characterized by physical discomfort (such as pricking, throbbing, or aching) and perceived to originate in the abdomen.	http://purl.obolibrary.org/obo/HP_0002027	Abdominal pain	An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.	http://purl.obolibrary.org/obo/HP_0012531	1
anxiety disorder	A category of psychiatric disorders which are characterized by anxious feelings or fear often accompanied by physical symptoms associated with anxiety.	http://purl.obolibrary.org/obo/MONDO_0005618	anxiety disorder	An anxiety disorder that is characterized by unexpected and repeated episodes of intense fear accompanied by physical symptoms	http://purl.obolibrary.org/obo/DOID_594	1

Between COVOC and CIDO there were (n= 26) same URI mappings with similarity scores above 0.95 that were included in the semantic matcher. There were (n=162) term

mappings with a similarity score of 1 after those with the same URI were removed. Terms with the same URI can be considered skos:exactMatch without review, however, the mappings with different URIs had to be manually reviewed. The results of the semantic matcher for term labels is shown in Table 10. Results between COVOC and COVID 19 are not reported in the table since only 19 mappings with scores above 0.95 were identified.

Table 10

Semantic Matcher Results

Term	**Similarity Score = 1**	**Similarity Score >0.95 and <1.0**
Covid-19 Vocabulary Ontology Coronavirus Infectious Disease Ontology **Same URI**	26	1
Covid-19 Vocabulary Ontology Coronavirus Infectious Disease Ontology **Same term label/different URI**	162	78
Coronavirus Infectious Disease Ontology Covid-19 Ontology **Same URI**	59	108

Coronavirus Infectious Disease Ontology Covid-19 Ontology **Same term label/different URI**	54	3

A review of the output mappings predicted in this workflow segment show that terms below with similarity scores below 0.99 and greater than 0.8 are highly likely to be related to each other. For example, the following:

Triazole antifungal agent ⇔ antifungal agent

Ribosomal large subunit assembly ⇔ ribosomal small subunit assembly

Immune system process ⇔ immune system disease / abnormality of immune system

These terms could be considered as broader or narrower terms, but need expert review to determine what the appropriate semantic type match is.

4.5 Output

At the end of these phases, combining and filtering of the results generated by one or more matchers was done. Therefore, we obtain first a set of mapped terms with their unique identifiers which can be accepted without oversight, the terms determined to be skos:exactMatch and skos:closeMatch. These are the terms used for the dictionary applied for tagging of clinical trials. In addition, a set of mapping suggestions which needs review by a domain expert who can then accept or reject them is also produced. Accepted suggestions could then be added to the final integrated clinical coding scheme.

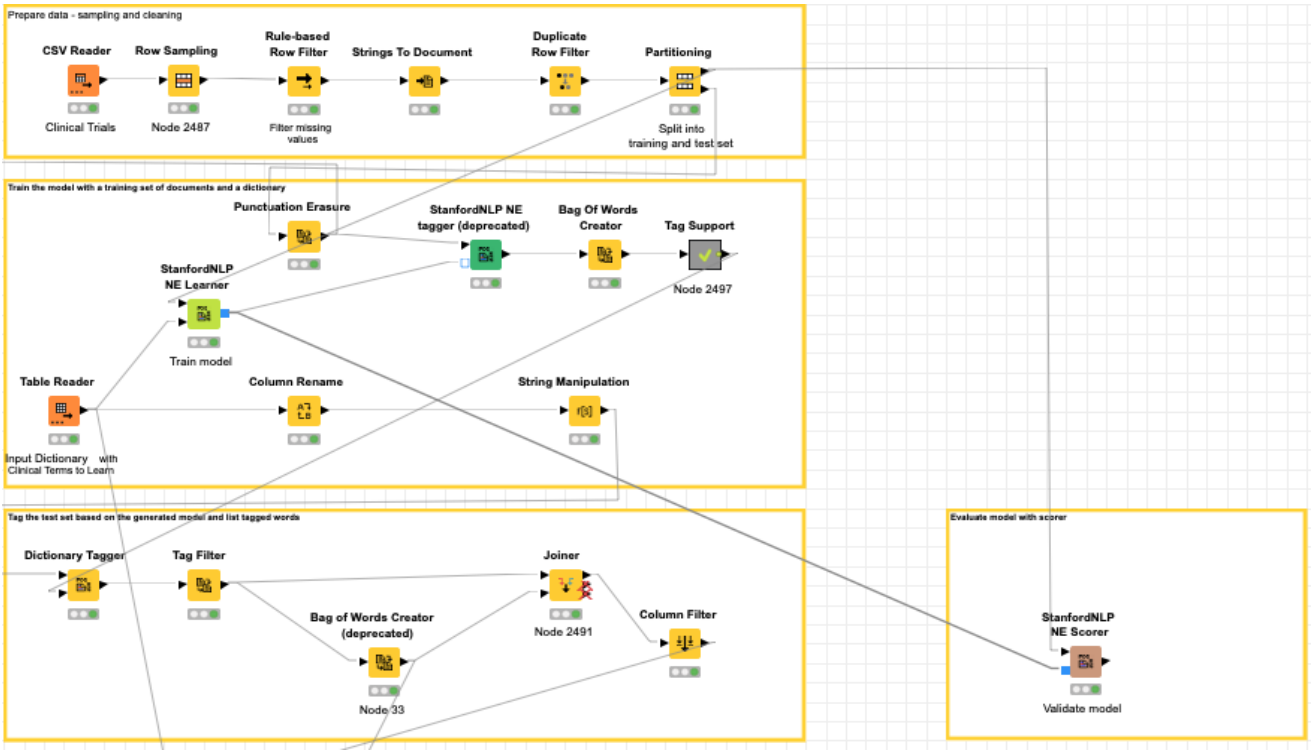
4.6 Clinical Trial Annotation Workflow

One objective of the project was to annotate clinical trial documents using the integrated set of terms obtained from the mapping workflow segments. To accomplish this, named entity

recognition (NER) a natural language processing (NLP) technique was implemented within the workflow. This workflow uses two sets of input data: the sample clinical trials a total of 575 documents and a set of dictionary terms and their unique resource identifiers. These dictionary terms are obtained from the combined output of the mapping workflow segments. Prior to beginning named entity tagging the data undergoes some preprocessing e.g., punctuation removal, and data type conversions. The data is then partitioned into test (30%) and training (70%) data for use in a conditional random field model applied with the Stanford NLP NE learner node (see Figure 21).

Figure 21

Section of Annotation Workflow



The conditional random field model created with the Stanford NLP NE Learner node used untagged sample clinical trial documents and the dictionary terms which are the named entities that should be identified in the documents to create a model which learns the dictionary terms and tries to correctly identify them in the clinical trial documents. The test data is then

passed to the NE tagger node which assigns named entity tags to the corpus of documents using the learned model that was trained on the dictionary of ontology terms. This node is later connected to another dictionary tagger to ensure the use of the terms specified in the dictionary and then to a tag filter to filter the terms in the input documents that have certain tags assigned to them. The results are then prepared for visualization through a series of join, group by, and viewer nodes. Figure 22 shows an example of some dictionary terms that were tagged in the documents.

Figure 22

Example of term tags present in dictionary and clinical trial documents

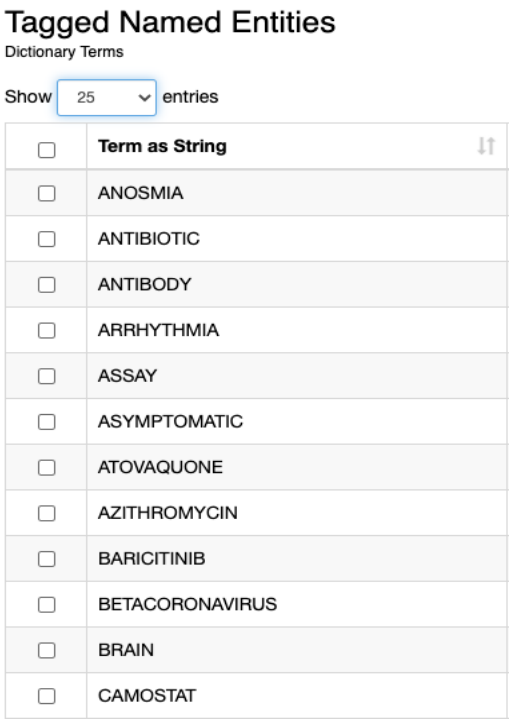
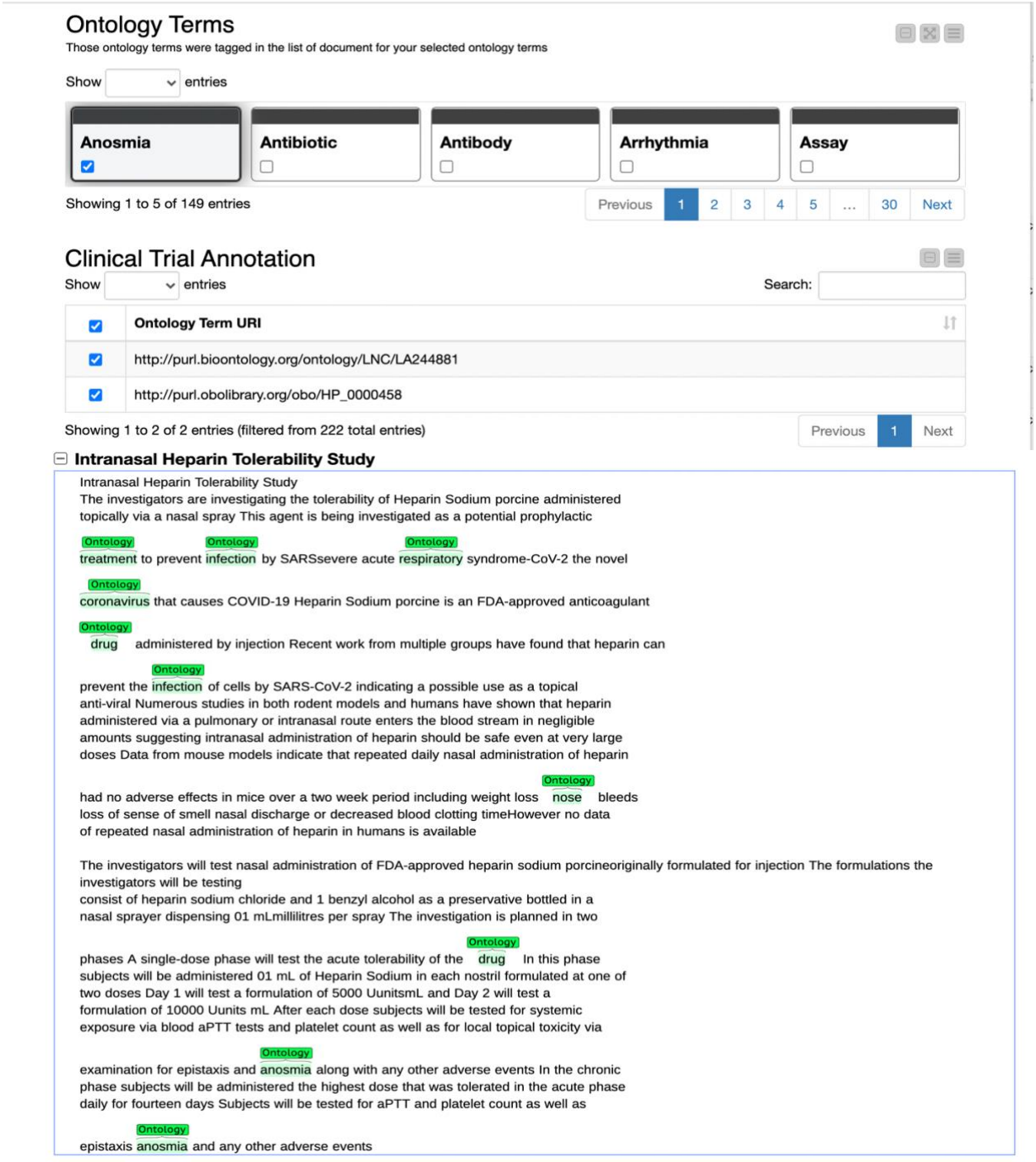


Figure 23

Example Output from Clinical Trial Annotation



Frequently occurring terms across documents include such terms as COVID-19, pneumonia, convalescent plasma, hydroxychloroquine, ivermectin, antibody, azithromycin,

hypoxia, and dyspnea among others. Some terms also tended to occur together frequently, for example ‘cancer’ and ‘treatment’, since several studies used patients undergoing treatment for cancer, examples of other cooccurring terms include hydroxychloroquinone and ritonavir, antibody and vaccination, isolation and quarantine or convalescent plasma and treatment. An example of the tagging document output for the test corpus is shown in Figure 23.

4.7 Evaluation Results

DSR research like other methods recognizes the need for evaluating knowledge outcomes and assessing the effectiveness and usefulness of artifacts that are produced (Larsen et al., 2020). In DSR research one recommended pattern for evaluation is Benchmarking. Benchmarking provides a vehicle for the objective evaluation of a solution or comparison of different solutions (Tichy, 1998) which makes it easy to verify that a claimed solution can solve a problem or is better than other existing solutions (Vaishnavi et al., 2015), in addition to suggesting a new method to address the problem.

The benchmark used for this research are a validation set of terms obtained from a subset of existing mappings within BioPortal and used as the Gold Standard against which to compare the workflow mapping results. The BioPortal mappings have been identified through either the NCBI’s LOOM algorithm, a UML unique concept identifier assigned by editors at the National Library of Medicine or through OBO referencing. These mapping results were obtained for the schemes of interest through the BioPortal REST API.

Three clinical coding schemes COVOC, CIDO and COVID-19 ontologies were used in the research. The COVOC ontology is not stored within BioPortal, and the repository in which it is stored does not display or make mappings available. However, the mapping results from the workflow between the remaining schemes can be checked and compared with the gold standard. Additionally, the lexical matcher was also run across data from LOINC primarily as an

additional means of testing the results obtained from the workflow. A list of the number and type of terms available from BioPortal Gold Standard set of data is provided in Table 11.

Table 11

Mappings Used for Validation

Mappings	SAME URI	LOOM	TOTAL
CIDO / COVID-19	587	79	666
CIDO/LOINC	0	871	871
COVID-19/LOINC	0	489	489

4.7.1 Lexical Matcher Metrics

The lexical series matcher identified 903 mappings between CIDO and COVID-19 ontology that were labeled as skos:exactMatch, skos:closeMatch or needing review. To evaluate the results, the mapping results were compared with the gold standard mappings. Of those, the matcher identified 602 correct matches between the CIDO and COVID-19 ontology, there were an additional 294 matches found that were not present in the Gold Standard. Between CIDO and LOINC the matcher identified 747 mappings of those 426 were correctly identified when compared with the gold standard mapping. There were an additional 252 matches found that were not present in the Gold Standard. Finally, between the COVID-19 ontology and LOINC, the matcher identified 398 mappings. There were 248 correctly identified terms from that set, however, there were 145 mappings found that are not in the Gold Standard. Performance scores were calculated for these algorithms using the Gold Standard set as the benchmark and shown in Table 12. An explanation for these measures was provided in the methodology section.

Table 12

Performance Measures as Calculated Based on Gold Standard

	Precision	Recall	F-measure
CIDO ⇔ COVID-19	0.979	0.926	0.961
CIDO ⇔ LOINC	0.861	0.477	0.681
COVID-19 ⇔ LOINC	0.996	0.511	0.767

Precision is high in most cases, even though recall is low in the tests with LOINC. When precision is high and recall is low, it is often because the algorithm is returning few results, but the predicted mappings are correct. This can occur when there is a class imbalance between the datasets. Additionally, this may be because additional data transformation is needed. Since the Covid-19 Vocabulary (COVOC) Ontology did not have a usable dataset that could function as a gold standard for which to compare the result of that workflow segment, however, out of 498 found matches with a distance score greater than 0.25, there were 146 correct matches, with an accuracy score of 0.293, calculated as correct matches divided by found matches. It was noted that in this particular match segment, matches greater than 0 were unlikely to be similar at all, even though in other segments, matches between 0 and 0.25 could not be immediately discarded since they contained some correct matches. This is probably due to differences in domain and coverage of ontology terms. Between COVOC and LOINC, there were 144 found matches and 121 of those were correct with an accuracy score of 0.840.

4.7.2 Document Similarity -Term Definition Matcher Metrics

In the Document Similarity -Term Definition Matcher, the accuracy measure is used to determine the performance of this workflow segment. For each scheme, only definitions with a

similarity score above 0.85 will be considered. Since this algorithm focuses on term definitions rather than term labels, the gold standard dataset is not used. Instead, the accuracy was assessed as determined by the number of correctly identified terms – those with a similarity score above 0.95 and which can be accepted without expert review, divided by the total number of predictions with a similarity score above 0.85.

These correct items were manually reviewed to ensure they fit the criteria. Between COVOC and CIDO the accuracy was calculated out of 111 total predictions at 0.234 with 26 correct predictions. Between COVOC and COVID-19 the accuracy was calculated out of 47 total predictions at 0.532 with 25 correct predictions. Finally, between CIDO and COVID-19, out of 186 found matches, the workflow has an accuracy of 0.467 with only 87 identified correct matches between CIDO and COVID-19. The structure of LOINC differed somewhat from these clinical coding schemes and did not include a consistent property for term definitions. Therefore, LOINC was not assessed with the document classifier.

One thing that can explain these results is the differences in number of items in each class. Accuracy scores tend to be low when there is class imbalance, however, this is one more way in which the clinical coding scheme might be enriched by considering an element of the ontologies that is not typically included. Another, explanation could be the intended coverage of the ontology. For example, the COVID-19 vocabulary ontology and the coronavirus infectious disease ontology while they both conceptualize a similar area, the CIDO is more generalized to all coronaviruses and their resultant diseases.

4.7.3 Semantic Matcher

The results of the semantic matcher workflow segments show reasonable results. This matcher compares every term against the entire set of terms in the other vocabulary and returns a score for each. It does this for every single term present in the source vocabulary. Due to processing constraints, only terms with similarity scores greater than 0.95 are considered for reporting. Correct mappings are defined as those with either the same URI or a semantic

similarity score of 1 that was manually verified. Among the semantic matches with different URIs and scores lower than 1, there also exist some matches. However, their usage cannot be automated without heuristic rules or through manual review to determine the type of relation that exists between the matches.

Table 13

Semantic Similarity Mappings

	Found Mappings	Correct Mappings	Accuracy
CIDO ⇔ COVID-19	314	206	0.656
COVOC ⇔ CIDO	369	291	0.789
COVOC ⇔ COVID-19	19	8	0.421

Accuracy results for the semantic similarity matcher (see Table 13) are moderate and is calculated as correct mappings divided by found mappings. Results from this matcher suggest that more heuristic rule-based methods would need to be included in the workflow to identify mappings with scores lower than that are related but not the same. The workflow segment for mappings between COVOC and the COVID-19 ontology produced very few mappings with scores above 0.85.

4.7.4 Clinical Trial Annotation Evaluation

To validate the clinical trial annotation workflow, a model scorer was implemented in the workflow. 90% of the obtained clinical trials was used and those with missing study descriptions were removed from the set leaving a total of 575 documents. The dataset was split into training and test data. The training set was comprised of 30% of the clinical trials drawn randomly from

the full sample, which resulted in 172 documents in the training set, with a remainder of 403 documents in the test set.

Calculation of the quality measures was done by passing the training set of data as well as the learned model to the NLP scorer node. This node tags the test set of data with a dictionary tagger internally based on the dictionary used for training. After the documents are tagged, the input model again tags the data and the differences between the tags created by the dictionary tagger and the input model are calculated. Table 14 shows the results of evaluation from the NLP model scorer.

Table 14

NLP Model Scorer Results

Precision	Recall	F1	TP	FP	FN
0.994	0.994	0.994	711	4	4

The model achieved high scores with 99% precision and 99% recall. These results indicate that the workflow for annotation is identifying more relevant results than irrelevant results and that most of the dictionary terms are identified, even if other terms are also identified.

4.7.5 Determining Functionality

A set of criteria for determining the functional level of the workflow was previously described and outlined as follows:

- If the tool produced half the number of similar matches, it would be considered partial functioning.
- An equal number of mappings would be considered similar or full functionality.

- If a greater number of mappings are found, the tool can be considered to meet or exceed the gold standard.

Based on the results of the various matchers, the workflow is demonstrably able to identify similar mappings to those present in the gold standard. For example, the lexical series matchers return more than half in all cases and almost all mappings in some cases as those in the Gold Standard. Additionally, it identifies some additional mappings with moderate to high precision and recall compared with the gold standard as shown in Table 12. Therefore, the lexical series matcher can be considered to have partial to full functioning capacity based on these results.

The semantic similarity matcher, however, can only claim partial functioning since most of the matches found with a score below 0.95 cannot be automatically accepted. For example, in the case of CIDO and COVID-19, only about half the number of terms in the gold standard are identified, and not all of these are true positives, i.e., skos:exactMatch or skos:closeMatch. Instead, this workflow segment seems ideal for identifying skos:relatedMatch types with either rule-based functions implemented or human oversight. This data could be presented to a clinical informatician for them to determine whether and what type of match exists.

The document annotation-term definition matcher produced the weakest results, however it compared term definition annotations rather than term label terms or meanings. A lack of definitions for many ontology terms had a negative impact on the quality of results, accuracy measures are low, but whether this is due to incorrect predictions or class imbalances is not clear. However, observations from this workflow about the definitions and structure of the vocabularies was helpful for providing insight into the process of scheme creation and the interaction between the human element and the system.

Finally, using the COVID-19 Vocabulary Ontology as a target, and primary focus of enrichment, when the results of all the matcher workflow segments was combined, there were a total of 450 mapped terms which amounts to 82% coverage in a scheme with 547 total terms, although it had fewer mappings to begin with and none to the clinical coding schemes tested in

this workflow. A final list of mapped terms across all vocabularies from all the workflows amounts to about 1395 mapped terms and their unique identifiers across vocabularies that were used as a dictionary of terms for the clinical trial annotations.

Chapter 5. Research Question 1 - How can an Extract Transform Load (ETL) workflow tool support the task of clinical coding scheme mapping?

5.1 Background

Semantic mapping has become important for enabling the translation of healthcare data between various types of core reference terminologies that support description of patient data, reporting, administrative or epidemiological classification and more. However, difficulties in accomplishing mapping and vocabulary enrichment such as granularity, structure, domain and language, data models, inconsistencies in concepts and meanings often mean complex and involved programmatic responses are put in place to accomplish it. However, ETL tools are relatively new or untried in the information organization space yet may support or constrain the ways in which mapping is performed across communities of practice. The lessons learned while attempting to create a workflow approach to mapping are outlined in this section.

5.2 Methods

Various strategies for developing a solution to a research problem through the creation of an artifact have been outlined for design science research (Hevner & Chatterjee, 2010a; Vaishnavi et al., 2015). This work used the Preliminaries Type pattern (Vaishnavi et al., 2015) where one goal of research is expanding the choice of tools and techniques that can be used to solve a research problem and determine whether a promising method is being overlooked and should be adopted by the research community. This is assessed by utilizing the KNIME workbench software to create a workflow for mapping and annotation of unstructured documents presented in chapter 4. The workflow uses functional nodes and connections and configurations between them to complete the tasks.

5.3 Findings

Since this work is partially about using a new tool to solve a problem, the findings are a description of the design process, the results obtained, and the evaluation of the processes involved. These have been reviewed in detail in Chapter 4 but in summary demonstrate that mapping and annotation can be accomplished through a workflow approach using ETL tools. While lexical series matcher returned the strongest results in terms of accuracy and alignment with the benchmarks. The semantic similarity matcher provided results that could be passed to a domain expert such as a clinical informatician to correctly label. However, the document similarity matcher while returning high precision scores, meaning its predictions were correct, suffered in terms of recall due to the lack of definitions for terms present across the clinical coding schemes tested. The ways in which the workflow artifact and by extension the ETL tool supports mapping are outlined in the discussion below.

5.4 Discussion

The findings described in chapter 4 show three methods for enabling mapping using a workflow in ETL workflow tools. There are opportunities for testing of different and more complex methods within the workflow in future iterations of the design process. However, the unique support that ETL workflow tools can offer for mapping are discussed in the next sections.

5.4.1 Facilitate easy loading and analysis of datasets

Loading data into these tools is simple. The tool offers a variety of input nodes ranging from csv, xml, json, table, to triple file readers. The workflow in this research used a triple file reader and SPARQL insert to load data into a in-memory endpoint. This simplifies data isolation and extraction as in this case where a SPARQL query is run and the results connected to data processing nodes. For example, .owl, .ttl files containing triples, often must be opened by tools such as Stanford's Protégé to view and assess their content and structure and determine what

the classes and properties of interest are, however ETL tools can provide this function natively, thereby cutting out the need for an extra step. In some cases, only API access to data is additionally available, such access can be enabled in the tool via the GET request node, removing the need for separately building out programmatic access to the web APIs.

5.4.2 Data Cleaning and Transformation

Often special software for data cleaning and transformation may be required to get the data into a special format before it can be used. However, ETL tools do not require data to be separately prepared as those capabilities are available within nodes that can be configured depending on the users' requirements. Therefore, data preparation simply becomes one of the initial steps in the workflow process

5.4.3 Reductions in operating cost

Mapping, and vocabulary enrichment are expensive and time-consuming processes. Simperl et al (2012) note that the development of clinical coding schemes is subject to a number of product, personnel, and project related cost drivers that can make or break the project. Product related complexities involved with domain analysis, conceptualization, implementation, instantiation, evaluation, documentation and required usability can be particularly impactful as these are all critical parts of the development process. Regarding personnel, the ontologist/domain expert capabilities and experience as well as language and tool experience, and the continuity of personnel also drives cost. Project-related cost drivers cover support tools for ontology engineering, multisite development, and the required development schedule (E. Simperl et al., 2012). Using an open source ETL tool may significantly lower the impacts of these cost drivers in the following ways.

5.4.3.1 Product, Personnel and Project Related Cost Reductions. Ontology methodologies often recommend reuse of classes and properties and creation of links to related concepts. Evaluation can be accomplished through comparison with another source or

configuring statistical and scoring nodes. The lexical series matcher demonstrates that this can be achieved efficiently and support conceptualization or evaluation. Multisite development may be achieved through sharing workflows in a community repository where collaborative work can occur. Developers can share their workflows, work together and update versions of their workflow and receive and provide feedback (Schmidt, 2021).

The requirements for reuse and sharing can be a significant cost drivers but must be a priority (B. Simperl et al., 2006). Criteria for increasing reuse are varied but ETL tools offer quick methods for parsing multiple schemes in a similar domain for concepts that can be reused, or which can be linked through semantic matches. The lexical semantic matcher workflow segment can make it easy to quickly find skos:exactMatch, skos:closeMatch and skos:relatedMatch type terms and give an indication of which terms/classes can simply be reused and which must be linked via some semantic relationship. Thus, the advantages here as summarized in this list as follows:

1. No software costs if open source ETL tools used
2. Support for review and comparison of concepts across ontologies
3. Support for domain analysis and conceptualization processes.
4. Easy identification and linking of similar concepts.
5. Quick evaluation of function.
6. Reductions in work and time spent identifying and labeling similar concept
7. Multisite development possibilities through shareable workflows

5.4.4 Supports Assessment and Improvement of Data Quality

5.4.4.1 Support for FIT Metric Impactful. The I3 FIT metric of maximizing the impact of a vocabulary through mapping with other vocabularies (Zeng & Clunis, 2020) is critical to creating Linked Open Data Knowledge Organization System products. Instead of siloed data, which does not interact with other data, this workflow can support institutions efforts to make

their data impactful, since it supports the enrichment of ontologies through the addition of mappings.

5.4.4.2 Support for FIT Metric Transformable. The workflow supports transforming data through quick assessment of whether a clinical coding scheme or other controlled vocabulary (both source and target) meets other FIT metrics such as T3, which recommends enabling extensibility through assessment of provenance via the presence or absence of certain properties such as skos:changeNote or prov:wasGeneratedBy, and T4 which recommends that a scheme support innovative and transformative uses beyond being normal value vocabularies. The workflow supports quick extraction of scheme properties, making it easy to check which are being used. This is accomplished through the data input node segments. Further with the implementation of clinical trial annotation, the workflow has transformed the purpose and use of the data beyond being an available vocabulary.

Since the tool allows the entire dataset or some subset of it to be easily integrated it is possible that it can be used in a variety of applications particularly those involving knowledge graphs. In addition, developers could easily build out and test queries within the tool as evidenced by those that were employed within the workflow, which can later be shared with the published dataset or provided as part of a workflow that allows exploration of the scheme. An example of these kinds of enrichment activities can be seen in the KNIME workflow FAIR data with KNIME which exemplifies how a workflow tool can make data FAIR and is published in the work by Delp et al (2018).

5.5 Conclusion

ETL tools support mapping by providing a simple interface in which mapping can be accomplished and evaluated. In addition, they offer several efficiencies such as reductions in operation costs and cognitive load, fast and easy deployment of solution, facilitation of

interoperability, easy maintenance and modification of schemes and insights into community-based development.

Chapter 6. Research Question 2 - How does the mapping output of the novel workflow support and affect annotation of clinical trials in COVID-19 research?

6.1 Background

The potential of secondary use of unstructured data available in medical documentation, clinical trials, and other text heavy clinical documents for improving patient care and outcomes through better diagnoses, treatments, and drug approval processes depends on the semantic annotation of unstructured data (Smithwick, 2015). The ability to provide structure to data by recognizing equivalent concepts and explicitly clarifying and adding consistence to the meaning of terms enhances the discoverability and usefulness of data for clinical professionals, researchers, and patients. With the emergence of the COVID-19 disease and its ongoing threat to humanity, many researchers have started or completed clinical trials, as well as other research activities, and additionally published their work.

This has led to an explosion in the amount of unstructured data that is available surrounding the topic. With this information glut comes a need to quickly label and identify the scope and content of the data. Semantic analysis provides methods and models for extracting information from unstructured data, crucially through the identification of named entities within the document. Semantic technologies involving machine learning, natural language processing, and pattern recognition are all useful for extracting knowledge from scientific data but recognizing named entities is the most critical step as it identifies terms or concepts (Zhu et al., 2013).

6.2 Methodology

For COVID-19 research, the identification of comorbidities, genes, cells, and other biological entities, as well as potentially applicable drugs and treatments is critical. Comparisons

between research studies, integration of data into a single context, inferencing, knowledge discovery, interpretation and reuse are challenging without aligning concepts in unstructured data to a KOS (Davies et al., 2006; Geraci et al., 1991; Gliklich et al., 2014; Smithwick, 2015). Clinical trials contain this sort of information therefore as outlined in Chapter 3 Named Entity Recognition (NER), an application of NLP, is implemented in the workflow to identify entities of interest in text blocks and add their unique ids from the clinical coding scheme of interest, through conditional random fields modeling. Model performance measures are assessed using available scoring nodes described in section 4.7.4 and the tagged documents are then manipulated and visualized.

6.3 Findings

The mapping segments of the workflow developed in this work resulted in a list of controlled terms along with their Unique Resource Identifiers (URIs) from various ontologies developed to deal with COVID-19 and coronaviruses. These terms are essential to linking entities of interest in the clinical trials with appropriate entries in the clinical coding schemes. This data formed the core of the dictionary used to train the NER model, that produced high precision and recall scores resulting in a F1 measure of 99% which is a single metric representing the harmonic mean of precision and recall. The resulting annotated data (see Figure 25) demonstrated appropriate tagging of the unstructured data with concepts from the dictionary. Although, there were terms that existed within the dictionary and by extension the various concept schemes, that may need filtering to reduce the noise of terms that are not specific to COVID-19 research.

6.3.1 Standard codes

During named entity tagging, the named entities contained in the dictionary are identified in the text along with the URIs. URIs in these vocabularies contain as a part of the name, the

standard code which distinguishes the concept. For example, the COVID-19 ontology term anosmia uses the standard code HP:0000458, this code is also a part of the URI which in full is http://purl.obolibrary.org/obo/HP_0000458. This holds true for most of the terms in the tested vocabularies. For example, see standard codes for LOINC as part of URIs in Figure 24.

Figure 24

Standard Codes as reflected in URI

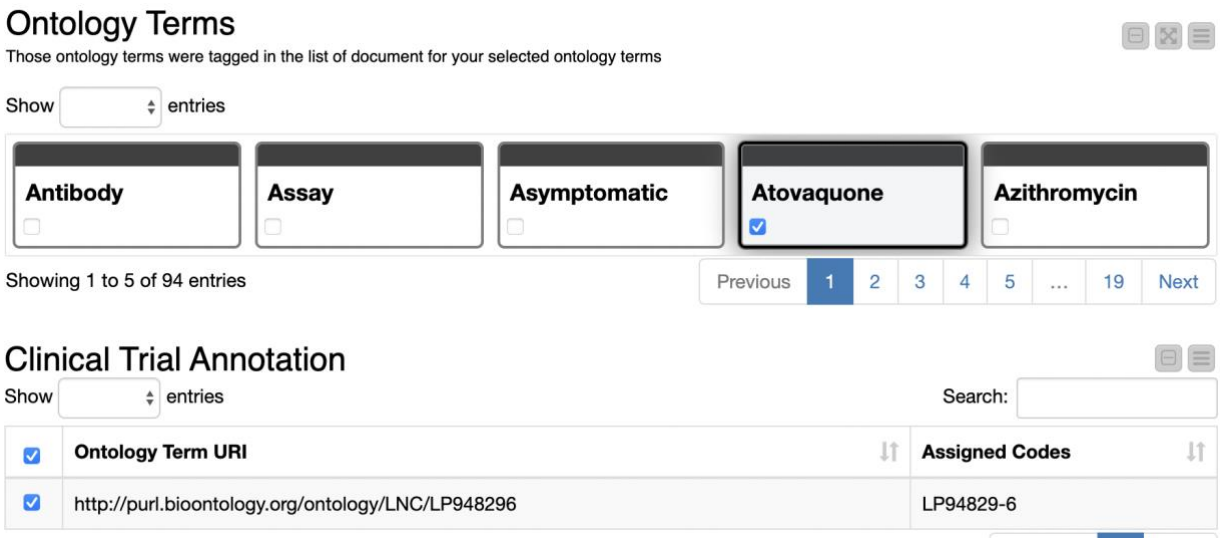
http://purl.bioontology.org/ontology/LNC/3880-2	http://www.w3.org/2004/02/skos/core#notation 3880-2
http://purl.bioontology.org/ontology/LNC/LP36416-3	http://www.w3.org/2004/02/skos/core#notation LP36416-3
http://purl.bioontology.org/ontology/LNC/23629-9	http://www.w3.org/2004/02/skos/core#notation 23629-9
http://purl.bioontology.org/ontology/LNC/LP396660-5	http://www.w3.org/2004/02/skos/core#notation LP396660-5
http://purl.bioontology.org/ontology/LNC/37728-3	http://www.w3.org/2004/02/skos/core#notation 37728-3
http://purl.bioontology.org/ontology/LNC/LP285195-6	http://www.w3.org/2004/02/skos/core#notation LP285195-6
http://purl.bioontology.org/ontology/LNC/52069-2	http://www.w3.org/2004/02/skos/core#notation 52069-2
http://purl.bioontology.org/ontology/LNC/78053-6	http://www.w3.org/2004/02/skos/core#notation 78053-6
http://purl.bioontology.org/ontology/LNC/LA13326-6	http://www.w3.org/2004/02/skos/core#notation LA13326-6
http://purl.bioontology.org/ontology/LNC/65917-7	http://www.w3.org/2004/02/skos/core#notation 65917-7
http://purl.bioontology.org/ontology/LNC/LP150222-0	http://www.w3.org/2004/02/skos/core#notation LP150222-0
http://purl.bioontology.org/ontology/LNC/MTHU059214	http://www.w3.org/2004/02/skos/core#notation MTHU059214
http://purl.bioontology.org/ontology/LNC/LP405540-8	http://www.w3.org/2004/02/skos/core#notation LP405540-8
http://purl.bioontology.org/ontology/LNC/85913-2	http://www.w3.org/2004/02/skos/core#notation 85913-2
http://purl.bioontology.org/ontology/LNC/4692-0	http://www.w3.org/2004/02/skos/core#notation 4692-0
http://purl.bioontology.org/ontology/LNC/LP229378-7	http://www.w3.org/2004/02/skos/core#notation LP229378-7
http://purl.bioontology.org/ontology/LNC/14996-3	http://www.w3.org/2004/02/skos/core#notation 14996-3
http://purl.bioontology.org/ontology/LNC/LP37228-1	http://www.w3.org/2004/02/skos/core#notation LP37228-1
http://purl.bioontology.org/ontology/LNC/LP374431-7	http://www.w3.org/2004/02/skos/core#notation LP374431-7

However, clinical coding schemes may additionally have other unique identifiers stored in a separate property. For example, KOS that are a part of the Unified Medical Language System (UMLS), may have Concept Unique Identifiers stored in the UMLS namespace e.g., the property *umls:cui.* LOINC has both concept unique identifier (CUI) and Terms and Semantic Type Identifier codes (TUI) (Shah et al., 2018) in the input dataset. In particular, Clinical coding schemes like the ones tested in the work may have these standard codes stored in the *oboinOwl:hasDbXref* property value space, e.g., UMLS:C0003126. These can be collected with

the data and passed to the annotation tool for tagging (See Figure 25) using a SPARQL Query or rule-based row filter node and join node.

Figure 25

Standard codes in annotation results



Tagging and enriching the clinical trial data with these named entities and standard codes enriches the quality and utility of the annotated text.

6.4 Discussion

Annotating entities provides semantic enrichment of words and additionally can benefit inference of the topic at large. A discussion of implications of the findings is presented in the following sections.

6.4.1 Support for Highly Specific Annotation Needs.

While there are publicly available, highly accurate pre-trained models for extraction of common entities, for example, person, location, organization, etc., certain applications require identification of more specific entities. Identifying concepts that are unique to the topic, makes it possible to perform intelligent knowledge extraction. The mapping output directly supports this

kind of annotation by providing a list of tailored terms that can be used to train the CRF model to identify the desired entity types within the unstructured text. The model is built around the training set and related names and can also be re-trained and fine-tuned with a new set of dictionary terms whenever there is a new clinical coding scheme of interest or when there are significant changes in the schemes that are being used. The workflow offers a way to quickly plug in these new schemes or a new dataset and quickly update the dictionary used for training the model, which would result in increased semantic enrichment of the unstructured clinical text.

6.4.2 Easily refine results

Another advantage offered is a way to quickly ascertain whether a clinical scheme is providing the type of annotations that will be considered ideal for a use context. Since the dictionary is built within and directly connected to the workflow, unplugging sources, and replacing them with another is a simple matter of changing input data or copying and pasting workflow segments to input new data. Additionally, building in rule-based filters to remove terms present in a clinical coding scheme, but which are not considered to be critical or particularly useful for annotation is simple.

For example, results of the clinical trial annotation show concepts such as disease, sars-cov-2, control, clinical, treatment, infection, patient, or public being identified as named entities. While these results are not incorrect, they are general and are unlikely to be terms that would be needed to filter documents of interest. Therefore, compiling a list of those terms and using a rule-based row filter node to remove them before passing dictionary terms to the CRF model is likely to increase the efficiency and impact of annotation. The advantage offered by the workflow tool is that this can be exactly tailored to the needs of the user.

6.4.3 Connect annotation to mapping tasks

Another factor to note is that annotation is connected to and is a natural extension of the novel workflow. There is no need for identification of another tool to perform any portion of the clinical

annotation workflow. Specifically, data preprocessing, dictionary creation, model building, training, and tagging are all embedded as part of the novel workflow. Visualization of results can be achieved by adding the required nodes, for example, a tag cloud to show term frequency distributions, or creation of a dashboard for exploring the annotated documents.

6.4.4 Extensible to other domains

The novel workflow and annotation segments is not limited to the field of biomedicine. Once required concept labels are extracted from the KOS of interest and used in the workflow, that mapped output data can be used as a dictionary of terms for annotation of other types of unstructured data. For example, in the pilot, collections of visual resources were used. The description of these images could be obtained as unstructured text and passed into the annotation workflow, a list of dictionary terms could be obtained by mapping KOS that provide standard terms for example the anthropology thesaurus or the art and architecture thesaurus. Other use cases could be to tag concepts in dissertation abstracts, or data from free text fields in electronic health records. Alternatively, the mapping portions of the workflow could be bypassed entirely, and a set of dictionary terms provided to the annotation workflow segment used to provide similar results. Basically, annotation of unstructured text can be tailored to KOS of interest without much extra work beyond accessing the data required.

6.5 Conclusion

The annotation workflow output demonstrated that the use of vocabulary terms enriched within the workflow with mappings from COVID-19 specific vocabularies offers the ability to provide rich indexing of clinical data for researchers to use or for downstream use in applications. In addition to the implications outlined in the discussion, annotation of unstructured documents also allows relationships to be made explicit through the hierarchical identification of concept labels and their corresponding classes. Further, legacy data can be quickly moved into the

future by adding semantic annotations which would allow the data to be searched and browsed. If an individual's or organization's goal is to quickly create a dictionary of terms within a specific domain for unstructured document annotation and to aid in knowledge discovery, then anyone with a fair understanding of the data being used in the project i.e., an informatics professional, data scientist, ontologist, researcher in any domain of interest, can adapt and use the workflow.

Chapter 7. Research Question 3 - What aspects of the sociotechnical model can be leveraged or updated to explain and assess mapping to achieve semantic interoperability in clinical coding schemes?

7.1 Background

In recent times, healthcare policy makers and stakeholders have emphasized the need for interoperable systems and mandated the adoption and use of clinical coding schemes in the information systems embedded in clinical care contexts. This convergence of human and technical factors creates a sociotechnical perspective from which clinical KOS should be considered.

7.2 Methodology

The reality of clinical coding schemes as knowledge organization systems suggest that they relate in various ways to the dimensions of the sociotechnical model developed by Sittig and Singh (2015). Therefore, their use in mapping, and the interaction between developers of clinical coding schemes, policy makers, technology, and stakeholder institutions can be viewed through this lens. Sittig and Singh's eight-dimension socio-technical model described in Section 3.4 above outlines the various dimensions. The social perspective is concerned primarily with the dimensions of people, workflow and communication, internal organization policies, procedures, and culture and External Rules, Regulations and Pressures. The technical perspective focuses on the Hardware and Software Infrastructure, Clinical Content, the Human Computer Interface, and System Measurement and Monitoring.

7.2.1 Theory in DSR

Previously (see section 3.4) it was stated that DSR knowledge contributions could include new theories or new knowledge that that serves to refine an existing theory (Lukka, 2003), partial or incomplete theory, or empirical generalizations from the research (Gregor &

Hevner, 2013). In design science research the phenomena of interest are created and so design theories can also include outcome specifications from which implications can be drawn. Vaishnavi and Kuechler (2015) mention that the design science knowledge usually starts out as an invention type of knowledge contribution and is accepted for the novelty and significance of the contribution and for the problem definition and solution/knowledge development standpoint. With the perspective of the artifact as an experimental apparatus the knowledge derived is what the design process can reveal about the complex socio-technical relationships behind the data input and systems. Therefore, observations across the socio-technical dimensions are informed by the literature and experiential knowledge gained while creating the novel workflow artifact.

7.3 Findings

The vision of interoperable systems that support large scale data sharing for increased quality in research and patient care faces many barriers. However, systems tend to reproduce the expectations, assumptions, and abstractions of designers and users (Ure et al., 2008). Since the COVID-19 pandemic began, international multi-center clinical trials and research teams, data sharing, and translational medicine applications have been developed as researchers and healthcare professionals seek try to mitigate the challenges of the pandemic. Clinical coding schemes support this vision by allowing reasoning across data sets of shared classes, properties, attributes, and relations representing a specific view of a domain.

7.3.1 Description of Scope

The clinical coding schemes in this research are all ontologies created to describe COVID-19. The number of class concepts in each ontology with the rdfs:label property varies greatly from a short list of 547 concepts in the COVID-19 Vocabulary ontology to an extensive 7866 terms covering not just the Sars-Cov-2 virus but all coronaviruses, in the Coronavirus Infectious Disease Ontology. In some cases, the ontologies reused classes and properties from

upper ontologies or other domain ontologies, but in other cases terms were unique. For each ontology the description of its scope was as follows:

> **COVID-19 ontology** – covers the role of molecular and cellular entities in virus-host-interactions, in the virus life cycle, as well as a wide spectrum of medical and epidemiological concepts
>
> **Coronavirus Infectious Disease Ontology (CIDO)** – provides standardized human- and computer-interpretable annotation and representation of various coronavirus infectious diseases, including their etiology, transmission, pathogenesis, diagnosis, prevention, and treatment
>
> **COVID19 Vocabulary Ontology (COVOC)** – covers terms related to the research of the COVID-19 pandemic. This includes host organisms, pathogenicity, gene and gene products, barrier gestures, treatments and more.

Based on these descriptions there was an expectation of more overlap than was found in terminology. For example, COVOC and CIDO should cover very similar concepts based on their descriptions, however only 146 exact or close matches were found with the lexical similarity algorithms and in the semantic similarity algorithms only 18 exact and 116 closely matching terms with an additional 383 needing review (see Table 8) were found.

7.3.2 Ontology Reuse and Linked Data

Another finding involved the reuse of concepts and linking of concepts to similar concepts. Although ontology reuse is important for knowledge representation, many concepts seemed to be created from scratch even though they may exist in another repository, or the concept seems to not exist elsewhere. For example, the terms below:

> http://purl.obolibrary.org/obo/COVOC_0030013 diammonium glycyrrhizinate
> http://purl.obolibrary.org/obo/NCIT_C102865 diammonium glycyrrhizinate
>
> https://bio.scai.fraunhofer.de/ontology/COVID_0000023 carriomycin

http://purl.obolibrary.org/obo/COVOC_0030010 carriomycin

Like these examples, sometimes there are concepts which are neither reused from another KOS, i.e., uses a URI from another scheme, nor does the ontology provide any links through properties such as skos:exactMatch, skos:closeMatch, or oboinOwl:hasDbXref, to other sources which may also use the term.

7.3.3 Data Governance

Another observation is that there is not always an indication of the concept source, in one case the fact that the term is obtained from a publication and the actual publication are provided in the concept class structure. In general, many times there was no indication of provenance or governance based on the properties used, for example no information about creators, contributors, or editors. In addition, there was a lack of definitions or descriptions for all concepts in the ontology.

7.4 Discussion

The discussion is organized according to social and technical dimensions and inform practical recommendations for improving clinical coding scheme development and the outcome of mapping projects.

7.4.1 Social Dimensions

People represents the stakeholders involved in the design, development, implementation and use of knowledge organization systems. In the context of this work the HIT in question is the clinical coding scheme. Workflow and Communication involved the identification of concepts and development of knowledge organization systems. Internal Organizational Policies, Procedures, and Culture and External Rules involve standards of practice for guiding and managing development and subsequent mapping of these schemes, as well as governmental, societal, and organizational pressures influencing these.

Clinical coding schemes are developed for and within specific communities of practice. For example, the coronavirus disease ontology (CIDO) describes itself as a community-based ontology which aims to provide an integration of the growth in data and research concerning COVID-19 and other coronaviruses (He et al., 2020). Although community engagement is critical to the success of a KOS, Ong and He (2016) suggest that opportunities for community involvement in ontology development is still limited despite them being created for community use.

7.4.1.1 People, Workflows and Communication. Successful community engagement requires asking and answering the question; who are the members of the community? This is important since clear identification and establishment of the community makes it easier to create best practices that govern the role and involvements of community members, in addition to the structure and function of the scheme itself as well as the scope of the concepts which will be included within it. Over the lifecycle of a knowledge organization system, a variety of people interact with it in different ways, the designers of the system, who could be researchers interested in developing an ontology solution to a problem or an organization creating a classification scheme to describe a field of interest such as the ICD which is used by public health officials for worldwide reporting, monitoring and comparison of health conditions, for health insurance billing and provider reimbursement, this means that it is built into health systems where clinicians, and patients can be impacted by its use.

If the KOS are to be mapped to each other or enriched to become linked open data, then domain experts are needed to ensure the correct interpretation of concepts and selection of equivalent terms. Based on the observations made about descriptions of scope, and concept development, greater community involvement of different kinds of people, for example ensuring ontology development project had individuals representing a variety of interests might ensure better quality datasets and metadata supporting the dataset that would reduce mapping challenges or make enrichment processes easier.

The presence of concept descriptions/annotation and verification of concepts within a scheme is an important feature for a terminology to adequately serve a community but is especially difficult to achieve in large schemes without the involvement of the community (Ong & He, 2016). In the experiment with document classification based on definition annotations, some classes were not included in the match workflow process since no definitions or other annotations were available for those terms. Figure 26 illustrates community members working together on concept development issues. In this example there is disagreement or uncertainty regarding the labeling of the term data set versus datum demonstrating differing opinions and beliefs at play in the people responsible for developing the clinical content.

This is representative of the workflow and communication dimension of clinical coding scheme development. The recommendation here is that properties are embedded within the schemas that allow editors to document the decision-making regarding concepts and structure. Community editing and discussion of these terms could be implemented through appropriate properties and with proper versioning support to allow for enrichment of the vocabulary. These changes would clarify concepts and make it easier to determine whether mappings are correct. Further, the use of tools that support data sharing and collaboration may impact mapping as well. For example, this novel workflow can be shared with the community at large, and their combined knowledge and experience may furnish ways to improve or modify it to give better results for their KOS of interest.

Figure 27

Community discussion of a term being conducted within a scheme

```
        <obo:IAO_0000116 xml:lang="en">2/2/2009 Alan and Bjoern discussing FACS run output
data. This is a data item because it is about the cell population. Each element records an
event and is typically further composed a set of measurment data items that record the
fluorescent intensity stimulated by one of the lasers.</obo:IAO_0000116>
        <obo:IAO_0000116 xml:lang="en">2009-03-16: data item deliberatly ambiguous: we merged
data set and datum to be one entity, not knowing how to define singular versus plural. So data
item is more general than datum.</obo:IAO_0000116>
        <obo:IAO_0000116 rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2009-03-16:
data item deliberatly ambiguous: we merged data set and datum to be one entity, not knowing how
to define singular versus plural. So data item is more general than datum.</obo:IAO_0000116>
        <obo:IAO_0000116 xml:lang="en">2009-03-16: removed datum as alternative term as datum
specifically refers to singular form, and is thus not an exact synonym.</obo:IAO_0000116>
        <obo:IAO_0000116 rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2009-03-16:
removed datum as alternative term as datum specifically refers to singular form, and is thus
not an exact synonym.</obo:IAO_0000116>
        <obo:IAO_0000116>2014-03-31: See discussion at
http://odontomachus.wordpress.com/2014/03/30/aboutness-objects-propositions/</obo:IAO_0000116>
        <obo:IAO_0000116 xml:lang="en">JAR: datum     -- well, this will be very tricky to
define, but maybe some
information-like stuff that might be put into a computer and that is
meant, by someone, to denote and/or to be interpreted by some
process... I would include lists, tables, sentences... I think I might
defer to Barry, or to Brian Cantwell Smith
```

7.4.1.2 Internal Organizational Policies, Procedures, and Culture and External Rules.

There is often conflict involved in the development of ontologies influenced by the Internal Organizational Policies, Procedures, and Culture and External Rules dimension of the sociotechnical model. Informed by the constraints of governmental policies and organizational needs, policies and procedures can be created to guide clinical scheme development. These can then be documented within the scheme through properties that support provenance or governance, and outside of the scheme itself in documentation documents.

Various conflicts can arise in the ontology development process itself or with mapping projects. Keet & Grutter (2021) outlines these conflicts as follows: *meaning negotiation* - which concerns deliberation to figure out the precise semantics that should be or which are represented in an ontology; *conflict resolution* - which concerns the choice among a set of two or more options; *language resolution* - conflicts where conflicts occur within a family of

languages or a distant one; and *ontological conflict resolution* - which involves philosophical decisions affecting the structure of the ontology or subject domain arguments with competing theories. Having firm policies in places, documentation included about the decisions made, the sources of data, the contributors, and editors of the coding scheme, makes it easier for those involved in either development, maintenance, mapping, or enrichment to perform the work they need to do.

Determinations on organizational budget may also limit software and human resource options for scheme mapping. Further, the intended usage context, rules and regulations may impact content and implementation decisions. The research artifact however demonstrates that it may be possible to create small scale solutions with minimal cost and within the context of organizational constraints, before moving to expand the scope of a mapping, enrichment, or new scheme development project.

7.4.1.3 Final Recommendations. Based on this discussion, final recommendations for questions to ask that will later have a positive impact on the quality of data produced and in turn on applications which use the data include:

1. Does selection and development of terminology involve stakeholders who will be likely to use it?
2. Are the terms specific to the context in which the clinical coding scheme will be deployed?
3. What governance structures were implemented while designing developing and implementing the clinical coding scheme?
4. What does governance structure mean in the context of the coding scheme being developed?
    a. Define approaches for identifying relationships between terms/classes
    b. Define processes for determining appropriateness of terms /collect feedback from potential users

c. Define processes for monitoring/maintaining the clinical coding scheme - versioning

7.4.2 Technical Dimensions

The technical dimensions address the hardware and software, clinical content, human computer interface and system measurement and monitoring procedures.

7.4.2.1 Clinical Content. The content dimension focuses standards, that is the knowledge organization systems or clinical coding schemes that support the interface between biomedicine and information technology. Use contexts, users, the data model acting as a framework for the classes, relationships and attributes, and the governance structures which support long-term viability are all critical components to consider when developing, using, and aligning data standards. Concepts (ontology class, term, property, and relationship labels) are social constructs, that is, they are ideas or perceptions of a thing based on the collective views developed and maintained within a society or social group who will have agreed that the concepts it exists and on the ways in which it may exist. By their very nature social constructs can change over time as they interact with the systems in which they are embedded. The design process revealed that features of the clinical coding schemes used, informed several recommendations for KOS development, mapping, and maintenance.

7.4.2.2 Software and Hardware. The software and hardware components used to store, create, and manipulate clinical coding schemes also have some impact on their utility. The size of the scheme has some influence on the hardware requirements of a system that will use it in terms of storage capacity and processing power. Applications in which the clinical content is deployed, need to be designed in ways that make best use of the structure, formats, and contents, rather than being strained or negatively impacted by them. For example, an out-of-date clinical coding scheme becomes problematic and less suitable to support the tasks it was designed to do, e.g., providing meaning, surveillance, data comparison, prediction,

discoverability, etc. Consideration should be given to the use of new software tools that can be used with clinical content for mapping, vocabulary enrichment and unstructured data annotation and other applications as necessary. ETL workflow tools are one such tool worth consideration as they can support semantic interoperability and the bridging of the gap between technical and human systems through supporting specific actions for enrichment of clinical coding schemes and other vocabularies.

7.4.2.3 System Measurement and Monitoring. KOS should be measured and monitored to determine their ongoing quality and suitability for functioning in complex systems. It is recommended that designers of KOS and those involved in mapping between them, consider the use of metrics to assess both the quality of the datasets (Wilkinson et al., 2016; Zeng & Clunis, 2020) and the mapping quality (Burrows et al., 2020; Randles et al., 2021).

7.4.2.4 Final Recommendations. The interplay of social construct embedded within information system in the form of the clinical content, provide unique opportunities for considering how clinical coding schemes support clinical applications, decision support, drug development, and research. From this mapping workflow experience, here are some general recommendations for clinical contents that can improve their utility in downstream applications.

- In a specific domain, reuse preferred labels and standard definition from more established schemes, rather than using a scheme specific label or variant definition. In the document annotation matcher, which reasoned over term definitions, variations in term definitions make it more difficult for machines to infer similarity.
- Provide as much enrichment as possible for concepts through reuse or providing semantic links – this means checking for terms in other KOS and reusing URI or if the context requires more specialized terms, and adding skos:mapping properties or oboInOwl:DbXref properites.
- Use preferred label properties for terms that are specific to a use context, but which may

exist in another form elsewhere, additionally consider using the alternative label properties to extend discoverability. The lexical series matcher provides quick mappings between similar terms based on lexical features that would support this process.

- Provide definitions for all concepts within a scheme - For terms in the scheme which lack definitions, review the definitions belonging to mapped terms and consider reusing them, or create a definition if none exists but add cross references to the term being defined, so that the definition will become discoverable.
- Use properties that indicate the source of concepts in addition to supplying information about term editors and contributors, especially in the context of newly discovered viruses and diseases. This will, in addition to facilitating collaboration, create opportunities for clarification of concept definitions or other special features of the concept, especially when that concept will be used in a scheme embedded in a new context.
- Use properties that document any conflict in concept description or labeling as well as those that support provenance and governance. CIDO for example integrates into its structure properties from the information artifact ontology to represent information such as editors, contributors, ongoing debates on term refinement. Providing this information makes it possible for content to be reviewed periodically with full knowledge of why certain decisions were made, who made them, and other such decisions related to maintenance and evaluation of the content within a clinical coding scheme.
- Review the ontology term labels in comparison to others to catch errors and anomalies, such as misspellings. Some errors might only be caught by those with local and contextual knowledge of terms when term labels are mapped have domain experts review those which cannot be automatically accepted, some term names are recorded differently across schemes, domain experts can verify whether these are the same terms or not.

- Provide options for community members to contribute to scheme development. Try to involve all categories of people who will interact with the scheme in term development, keeping in mind that content decisions vary across people groups as their understanding of concepts vary to match their realities. Terms should be as close as possible to those likely to be used by people in the domain in question. The semantic matcher workflow segment for example can act as a focus for debate on which concept is appropriate to include in a coding scheme.
- Make collaborative decisions on how the content will be made available in terms of formats, availability for download, API access, mapping schedules, software and so forth. These decisions determine how and whether the clinical coding scheme is used or reused or leveraged for research. Inaccessible content is a barrier to interoperability.
- Use standard KOS particularly mapped and enriched schemes to annotate unstructured data that cover similar topics rather than relying on pretrained models. This offers the advantage of tagging resources in a way that provides efficiencies for researchers within a certain domain.
- Explore non-traditional tools or methods for problem solutions. While these new software components may not be ideal for development of clinical content, they offer options for manipulation of the content that may offer new insights and efficiencies for the systems currently in place.

7.5 Conclusion

Clinical coding schemes may at first glance seem to only be addressed by the clinical content dimension of the socio-technical model. However, their nature as systems, knowledge organization systems, mean that the entire eight dimensions of the model can be used to address their development and use in applications and within other systems. Throughout the design of the novel workflow, observations and inferences were made about the

sociotechnical nature of clinical coding schemes which has led to several recommendations for improving the quality of clinical content and their utility in applications.

Chapter 8. Synthesis and Summarization

8.1 Background Restatement

Clinical coding schemes represent the underlying structure of a domain (Zeng et al., 2020). They express diseases, diagnoses, treatments, findings, operations, observations, medications, administrative and research concepts and more in the clinical domain (OpenClinical, 2005). A growing number of such schemes in addition to a lack of consistent usage across applications impedes data sharing and aggregation, increases communication difficulties, and creates challenges in the systems that depend on them. Schriml (2020) notes the critical nature of clinical coding schemes as infrastructure that support the proper functioning of healthcare systems and for facilitating data-driven research discoveries.

Problems caused by a lack of mapped data, semantic harmonization and terminology integration can blunt researchers' ability to perform the important and often lifesaving work they must do. The introduction and literature review established the difficulty of mapping and the various complicated methods used to perform the work. This research introduced and confirmed the use of new tools for addressing the problem of mapping and for supporting the annotation of unstructured clinical trial data. Euzenat and Shavaiko's (2013) classifies matching approaches as those based on either the element level or the structure level of the scheme.

This work addresses only the semantic and syntactic factors of the element level using terminological, syntactic and semantic matching techniques. Saitwal (2012) recommends that regardless of the method applied for mapping, it should be one that is easily reproducible when source terminologies are updated. Using the ETL tool to design a workflow for mapping and annotation of data, demonstrated that alternate methods for mapping are viable. Developing mappings across clinical coding schemes representing complex domains can be approached from a perspective which involves minimal operation costs and cognitive load, fast deployment of solutions, facilitation of interoperability, easy maintenance, and modification of schemes, as

well as insights into community-based development that approach the challenge of mapping as a collaborative task.

8.2 Summary of Outcomes

The design of the workflow artifact within the ETL tool showcases the nodes and configurations which can enable mappings between concepts and demonstrates how some algorithms and methods described in the literature review for mappings might be implemented in an ETL tool without the need of programmatic interfaces to establish them. It further uses the results of those mapping workflow segments to collect a set of terms and URIs that form a dictionary which train a ML model to annotate unstructured text implemented within the artifact and as an extension of the mapping workflows.

The approach was evaluated through benchmarking and comparison of the results with a gold standard set of terms. The workflow artefact was determined to perform at a similar standard as other standard tools with some workflow segments achieving only partial functioning with a potential for improvement with refinements. Review of the research from a sociotechnical lens led to a view of the clinical contents as more than simply a dimension of the model, but as knowledge organization systems whose development and usage can be viewed through the lens of all eight dimensions. The clinical coding scheme as envisioned as a social construct influenced by the social and technical dimensions of people, internal and external socio-political factors, software and hardware, and clinical content as concepts. A specific set of recommendations for content development based on observations made in the artifact design process was outlined.

8.3 Implications and Recommendations

A major tenet of the semantic web, FAIR data, and interoperability is the principle of reuse of resources. Secondary use of data lost in unstructured portions of clinical text can provide insights leading to applications and solutions that improve patient safety and health outcomes, reduce medical errors, enhance discoverability of cutting-edge research, and enable decision support. Although much work has been done in information extraction focusing on scientific literature, not as much has been done where NLP has been used to curate clinical trial fields (Miftahutdinov et al., 2021).

The workflow created in the ETL tool uses mapped terms to create a dictionary that is utilized for named entity recognition. What the tool contributes here is a simple method for researchers to annotate these documents making it easier to discover resources that are of interest. Further, the ability to annotate using a KOS of choice, or to integrate multiples of those enhances the quality of the annotations provided. Another output of the research is a shareable workflow artifact that can be reproduced as needed, adapted to suit special needs, or embedded in a more complex application. Although some mapping tools are available to the public for use, they do not allow the users the freedom to adapt their functioning in ways that more closely suit the project needs. Using an ETL tool, gives the freedom to test a variety of mapping methods including those that are current standard, but also allow the possibility of including and testing more modern approaches for any KOS of interest.

A wide variety of data is accessible with ETL tools, however, working with triples is made significantly easier than other methods, such as manipulating graph databases. This means that new ontologies and new data can be immediately accessed, loaded, mapped and/or enriched with minimal complexity. Even in the ontology development process, being able to identify similar concepts in other vocabularies is helpful for making data FAIR or Functional, Impactful and Transformable (FIT) (Zeng & Clunis, 2020). For example, the Gender, Sex, and Sexual Orientation (GSSO) Ontology currently reports beta status in BioPortal, however similar

terminologies exist outside the biomedical domain, including the Homosaurus LGBTQ+ linked data vocabulary, designers could utilize this tool to enrich the datasets before releasing a subsequent version, this would in turn increase findability of the resource not included within the BioPortal repository.

In previous chapters various efficiencies and advantages were discussed and in summary include rich indexing of clinical data for researchers to use or for downstream use in applications, explicit definition of relationships through hierarchical identification of concept labels and their corresponding classes, reductions in operation costs and cognitive load, fast and easy deployment of solution, facilitation of interoperability, easy maintenance and modification of schemes and insights into community-based development. Specific recommendations can be found in previous chapters.

Using an ETL tool is an easy way to update/make ontologies interoperable by identifying matches to terms that can be added through SKOS relationships or through classes which can be reused. It requires no programming or startup costs, is user friendly and only requires familiarity with the techniques or willingness to explore and learn how to implement them through the available nodes. Workflows created in the tool are expandable and configurable to specific context and can be adapted as needed. The ability to quickly enrich new schemes in high stakes contexts such as responding to the demands of a worldwide pandemic, is one which should not be overlooked by researchers in this space.

8.4 Limitations

This study was limited to only four clinical coding schemes, three dealing with COVID-19 and one tool mandated by government and used for lab reporting for benchmarking. Since, clinical trials often have much data related to drug testing, terminologies such as RxNorm could add benefit but were not included in this iteration. This is marked for inclusion in future versions of the workflow. This work focuses on the terminological features of the clinical coding schemes.

While solutions that focus on these features usually return high accuracy the recall results can often be low due to complexities with variations in the form of terms or labels. Therefore, future development should include methods that make use of external knowledge bases, or unsupervised and representation learning. Currently, the dictionary lacks entity type information such as whether a recognized entity is a drug, disease, cell, gene, et cetera. While not detrimental to functioning, adding this kind of data adds richness to the results in addition to enriching with URIs.

Mappings based on semantics can require significant processing power to compute. Certain refinements were computationally expensive, and if being done on personal machine should be accounted for. Another limitation has to do with the types of mappings or relationships between schemes. Mapping with high similarity scores and similar URIs have been labeled as either skos:exactMatch or skos:closeMatch, however, sometimes the concepts in an ontology have been imported from another ontology, rather than being a similarly named created concept. Additionally, the work did not use human experts to verify the data but instead used benchmarking with a gold standard alone.

8.5 Future Work

One interesting direction for future work would be to use the results from these matchers and pass them to more complex machine learning algorithms such as the association rules or support vector machines models. Current work in mapping and document annotation is testing deep learning models and other unsupervised methods (Chakraborty et al., 2021; Chen et al., 2020; Dhayne et al., 2021; Wang et al., 2021; Yan et al., 2021). Nodes to enable some of these are available as part of ETL tools and some of these for example building models based on word2vec, are methods that should be implemented and tested.

Another direction to explore is building in rules to automatically determine narrower and broader matches based on matches that use concepts labels are more or less specific (Zhou et

al., 2012). Further development of the annotation portion of the workflow is also another goal, refining the model and utilizing a larger dataset of unstructured terms might offer refinements that are not currently realized. Finally, using upper-level ontologies as a basis for structure-based mapping approaches is another direction that is worth exploring, these are helpful to mappings when adjacent elements are similar as structure-based matchers use taxonomy hierarchy or property attributes for processing.

Appendix A

List of Abbreviations

Abbreviation	Meaning
AAT	Art and Architecture Thesaurus
AHIMA	American Health Information Management Association
API	Application Programming Interface
CIDO	Coronavirus Infectious Disease Ontology
COVID-19	Coronavirus Disease 2019
COVOC	Coronavirus Vocabulary
CPOE	Computerized Provider Order Entry
CRF	Conditional Random Fields
CUI	Concept Unique Identifier
DSR	Design Science Research
EHR	Electronic Health Record
ETL	Extract Transform Load
FAIR	Findable, Accessibility, Interoperability, Reuse
FAST	Faceted Application of Subject Terminology
FIT	Functional, Impactful, Transformable
FDA	Federal Drug Administration
HIMSS	Health Information Management Systems Society
HIT	Health Information Technology
HTTP	Hypertext Transfer Protocol
ICD	International Classification of Diseases
ISO	International Standards Organization
KOS	Knowledge Organization System

LCSH	Library of Congress Subject Headings
LOD KOS	Linked Open Data Knowledge Organization Systems
LOINC	Logical Observation Identifiers Names and Codes
MeSH	Medical Subject Headings
NER	Named Entity Recognition
NLP	Natural Language Processing
RDFS	Resource Description Framework Schema
SKOS	Simple Knowledge Organization Systems
SNOMED CT	Systemized Nomenclature of Medicine Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
WHO	World Health Organization

References

AHIMA and AMIA Terminology and Classification Policy Task Force. (2009, July 22). *Healthcare terminologies and classifications: An action agenda for the United States.* Perspectives. https://perspectives.ahima.org/healthcare-terminologies-and-classifications-an-action-agenda-for-the-united-states/

Aho, A. V., & Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, *18*(6), 333–340. https://doi.org/10/crw9b4

Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). The social space of metadata: Perspectives of LIS academics and postgraduates on standards-based and socially constructed metadata approaches. *Journal of Library Metadata*, *4*, 311.

Allones, J. L., Martinez, D., & Taboada, M. (2014). Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. *Journal of Medical Systems*, *38*(10), 134. https://doi.org/10/f6hzxj

Antidote. (2021). *Patient Engagement during COVID-19: An analysis before and during the pandemic, with a focus on clinical trial patient recruitment*. https://www.antidote.me/antidote-whitepapers

Arvanitis, T. (2014). Semantic interoperability in healthcare. In J. Mantas, M. Househ, & A. Hasman (Eds.), *Integrating Information Technology and Management for Quality of Care* (p. 5). IOS Press.

Barrows, R. C., Cimino, J. J., & Clayton, P. D. (1994). Mapping clinically useful terminology to a controlled medical vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 211–215.

Bekhuis, T., Demner-Fushman, D., & Crowley, R. S. (2013). Comparative effectiveness research designs: An analysis of terms and coverage in Medical Subject Headings (MeSH) and Emtree. *Journal of the Medical Library Association*, *101*(2), 92–100. https://doi.org/10/f4zzsx

Bellahsene, Z., Emonet, V., Ngo, D. H., & Todorov, K. (2017). YAM++ Online: A web platform for ontology and thesaurus matching and mapping validation. In *ESWC: European Semantic Web Conference*, *LNCS*, 137–142. https://doi.org/10/ggbhp2

Binding, C., & Tudhope, D. (2016). Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, *17*(1), 5–21. https://doi.org/10/gf5vjv

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267–D270. https://doi.org/10/bzs9ps

Bodenreider, O., Cornet, R., & Vreeman, D. J. (2018). Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. *Yearbook of Medical Informatics*, *27*(1), 129–139. https://doi.org/10/gf33j8

Burrows, E. K., Razzaghi, H., Utidjian, L., & Bailey, L. C. (2020). Standardizing clinical diagnoses: Evaluating alternate terminology selection. In *AMIA Summits on Translational Science Proceedings*, *2020*, 71–79.

Carayon, P. (2006). Human factors of complex sociotechnical systems. *Applied Ergonomics*, *37*(4), 525–535. https://doi.org/10/dgv2cq

Carayon, P., Bass, E. J., Bellandi, T., Gurses, A. P., Hallbeck, M. S., & Mollo, V. (2011). Sociotechnical systems analysis in health care: A research agenda. In *IIE Transactions on Healthcare Systems Engineering*, *1*(3), 145–160. https://doi.org/10/fq87gc

Chakraborty, J., Bansal, S. K., Virgili, L., Konar, K., & Yaman, B. (2021). OntoConnect: Unsupervised ontology alignment with recursive neural network. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 1874–1882. https://doi.org/10.1145/3412841.3442059

Chen, X., Xie, H., Cheng, G., Poon, L. K. M., Leng, M., & Wang, F. L. (2020). Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences*, *10*(6), 2157. https://doi.org/10.3390/app10062157

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10/bs7c4p

Chute, C. G. (2000). Clinical classification and terminology. *Journal of the American Medical Informatics Association*, *7*(3), 298–303.

Chute, C. G., Elkin, P. L., Sherertz, D. D., & Tuttle, M. S. (1999). Desiderata for a clinical terminology server. In *Proceedings of the AMIA Symposium*, 42–46.

Cimino, James. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, *37*(4–5), 394–403.

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A Comparison of string distance metrics for name-matching tasks. In *IIWeb*, *2003*, 73–78.

Colic, N., Furrer, L., & Rinaldi, F. (2020). *Annotating the pandemic: Named entity recognition and normalisation in COVID-19 literature*. https://openreview.net/forum?id=QbCLrKBvurm

Davies, J., Studer, R., & Warren, P. (2006). *Semantic web technologies: Trends and research in ontology-based systems*. John Wiley & Sons.

De Quiros, F. G. B., Otero, C., & Luna, D. (2018). Terminology services: Standard terminologies to control health vocabulary: Experience at the Hospital Italiano de Buenos Aires. *Yearbook of Medical Informatics*, *27*(01), 227–233. https://doi.org/10.1055/s-0038-1641200

Delp, J., Gutbier, S., Klima, S., Hoelting, L., Pinto-Gil, K., Hsieh, J.-H., Aichem, M., Klein, K., Schreiber, F., Tice, R. R., Pastor, M., Behl, M., & Leist, M. (2018). A high-throughput approach to identify specific neurotoxicants / developmental toxicants in human neuronal cell function assays. *ALTEX - Alternatives to Animal Experimentation*, *35*(2), 235–253. https://doi.org/10/gnjfqz

Dhayne, H., Kilany, R., Haque, R., & Taher, Y. (2021). EMR2vec: Bridging the gap between patient data and clinical trial. *Computers & Industrial Engineering*, *156*, 107236.

https://doi.org/10.1016/j.cie.2021.107236

Dias, T., Alves, D., & Felipe, J. (2014, October 15). Method for the mapping between health terminologies aiming systems interoperability. In *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services*, *Healthcom 2014*. https://doi.org/10/ggbkn5

Edwards, P. N. (2004). “A vast machine”: Standards as social technology. *Science*, *304*(5672), 827–828. https://doi.org/10/c2pb67

EMBL-EBI Ontology Lookup Service. (2021). *CoVoc coronavirus vocabulary < Ontology lookup service < EMBL-EBI*. OLS Ontology Search. https://www.ebi.ac.uk/ols/ontologies/covoc

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching* (2nd ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38721-0

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10/gb345p

Fung, K. W., Xu, J., Rosenbloom, S. T., & Campbell, J. R. (2019). Using SNOMED CT-encoded problems to improve ICD-10-CM coding-A randomized controlled experiment. *International Journal of Medical Informatics*, *126*, 19–25. https://doi.org/10/ggmd7t

Gaudet-Blavignac, C., Foufi, V., Bjelogrlic, M., & Lovis, C. (2021). Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: Systematic scoping review. *Journal of Medical Internet Research*, *23*(1), e24594. https://doi.org/10/ghzxsj

Geraci, A., Katki, F., McMonegal, L., Meyer, B., Lane, J., Wilson, P., Radatz, J., Yee, M., Porteous, H., & Springsteel, F. (1991). *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries*. IEEE Press.

Giunchiglia, F., Shvaiko, P., & Yatskevich, M. (2004). S-Match: An algorithm and an implementation of semantic matching. In C. J. Bussler, J. Davies, D. Fensel, & R. Studer

(Eds.), *The Semantic Web: Research and Applications* (pp. 61–75). Springer. https://doi.org/10/b4t35k

Gliklich, R. E., Dreyer, N. A., & Leavy, M. B. (2014). Data elements for registries. In *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition*. Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK208639/

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*, 337–356. https://doi.org/10/gd3vw6

He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., Huang, H., Beverley, J., Hur, J., Yang, X., Chen, L., Omenn, G. S., Athey, B., & Smith, B. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific Data*, *7*(1), Article 1. https://doi.org/10/ghrsn6

Henricksen, K., & Kaye, R. (2003). Industrial ergonomic factors in the radiation oncology therapy environment. In K. Jorgensen & R. Nielson (Eds.), *Advances In Industrial Ergonomics And Safety V* (p. 325). Taylor & Francis.

Hevner, A., & Chatterjee, S. (2010a). Design science research frameworks. In A. Hevner & S. Chatterjee (Eds.), *Design Research in Information Systems: Theory and Practice* (pp. 23–31). Springer US. https://doi.org/10.1007/978-1-4419-5653-8_3

Hevner, A., & Chatterjee, S. (2010b). Design science research in information systems. In A. Hevner & S. Chatterjee (Eds.), *Design Research in Information Systems: Theory and Practice* (pp. 9–22). Springer US. https://doi.org/10.1007/978-1-4419-5653-8_2

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hochheiser, H., Ning, Y., Hernandez, A., Horn, J. R., Jacobson, R., & Boyce, R. D. (2016). Using nonexperts for annotating pharmacokinetic drug-drug interaction mentions in product labeling: A feasibility study. *JMIR Research Protocols*, *5*(2), e40. https://doi.org/10/ggktws

Hussain, S., Sun, H., Sinaci, A. A., Laleci, G., Mead, C. N., Gray, A. J. G., McGuinness, D. L., Prud'hommeaux, E., Bozec, C. D.-L., & Forsberg, K. (2014). A framework for evaluating and utilizing medical terminology mappings. *Studies in Health Technology and Informatics*, *205*, 594–598. https://doi.org/10/ggh8v7

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

In, J. (2017). Introduction of a pilot study. *Korean Journal of Anesthesiology*, *70*(6), 601–605. https://doi.org/10/gcn76b

Institute of Medicine (US) Committee on Data Standards for Patient Safety, Aspden, P., Corrigan, J. M., Wolcott, J., & Erickson, S. M. (2004). Health care data standards. In *Patient Safety: Achieving a New Standard for Care*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK216088/

ISO. (2018). *ISO 17117-1:2018(en), Health informatics—terminological resources—part 1: characteristics* (Standard ISO 17117-1:2018(en)). International Standards Organization. https://www.iso.org/obp/ui/#iso:std:iso:17117:-1:ed-1:v1:en

Jones, M., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Stevan, E., & Chong, S. (2019). Ecological metadata language version 2.2.0. *KNB Data Repository*. https://doi.org/10.5063/F11834T2

Juckett, D. (2012). A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics*, *45*(3), 460–470. https://doi.org/10/fzg3n9

Junker, M., Hoch, R., & Dengel, A. (1999). On the evaluation of document analysis components by recall, precision, and accuracy. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, 713–716. https://doi.org/10/b8k2mq

Kalra, D., Musen, M., Smith, B., Ceusters, W., & De Moor, G. (2011). ARGOS policy brief on semantic interoperability. *Studies in Health Technology and Informatics*, *170*, 1–15.

Kasanen, E., Lukka, K., & Siitonen, A. (1993). The constructive approach in management accounting research. *Journal of Management Accounting Research*, *Fall*, 23.

Katsumi, M., Gr&#252, & Ninger, M. (2016). What Is ontology reuse? *Formal Ontology in Information Systems*, 9–22. https://doi.org/10.3233/978-1-61499-660-6-9

Keet, C. M., & Grütter, R. (2021). Toward a systematic conflict resolution framework for ontologies. *Journal of Biomedical Semantics*, *12*(1), 15. https://doi.org/10/gnjfts

Kolyvakis, P., Kalousis, A., Smith, B., & Kiritsis, D. (2018). Biomedical ontology alignment: An approach based on representation learning. *Journal of Biomedical Semantics*, *9*(1), 21. https://doi.org/10/ggbhp4

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems*, *17*(5), 489–504. http://dx.doi.org/10.1057/ejis.2008.40

Kuperman, G. J. (2011). Health-information exchange: Why are we doing it, and what are we doing? *Journal of the American Medical Informatics Association*, *18*(5), 678–682. https://doi.org/10/dqcg34

Larsen, K., Lukyanenko, R., Mueller, R., Storey, V., Vander Meer, D., Parsons, J., & Hovorka, D. (2020). *Validity in design science research*.

Liang, S., Porat, T., Tapuria, A., Ethier, J.-F., Delaney, B. C., & Curcin, V. (2016). A dynamic medical terminology mapping system – MeTMapS. *A Dynamic Medical Terminology Mapping System-MeTMapS*, 7.

Lukka, K. (2003). The constructive research approach. In *Case Study Research in Logistics* (pp. 83–101).

McCulloch, E., Shiri, A., & Nicholson, D. (2005). Challenges and issues in terminology mapping: A digital library perspective. *The Electronic Library*, *23*(6), 671–677. https://doi.org/10.1108/02640470510635755

Menachemi, N., Rahurkar, S., Harle, C. A., & Vest, J. R. (2018). The benefits of health

information exchange: An updated systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1259–1265. https://doi.org/10/gdggjc

Miftahutdinov, Z., Kadurin, A., Kudrin, R., & Tutubalina, E. (2021). Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, *37*(21), 3856–3864. https://doi.org/10.1093/bioinformatics/btab474

Miles, A., & Bechhofer, S. (Eds.). (2009). *SKOS Simple Knowledge Organization System Reference*. https://www.w3.org/TR/skos-reference/#semantic-relations

Monge, A., & Elkan, C. (1997). *An efficient domain-independent algorithm for detecting approximately duplicate database records*.

National Center for Biomedical Ontology. (2021). *NCBO BioPortal*. https://bioportal.bioontology.org/

National Institutes of Health. (2017). *NIH's definition of a clinical trial*. NIH.Gov. https://nihodoercomm.az1.qualtrics.com/jfe/form/SV_eypqaXlx2j1IY9T?Q_CHL=si&CurrentPageURL=https%3A%2F%2Fgrants.nih.gov%2Fpolicy%2Fclinical-trials%2Fdefinition.htm&PageReferrer=null&Intercept=HTML&Q_CanScreenCapture=1

National Library of Medicine. (2021). *Glossary of common site terms—ClinicalTrials.gov*. ClinicalTrials.Gov. https://clinicaltrials.gov/ct2/about-studies/glossary

National Library of Medicine. (2019). *Introduction to MeSH* [technical documentation]. National Library of Medicine; U.S. National Library of Medicine. https://www.nlm.nih.gov/mesh/introduction.html

National Library of Medicine. (2020). *RxNorm overview [product, program, and project descriptions]*. U.S. National Library of Medicine. https://www.nlm.nih.gov/research/umls/rxnorm/overview.html

Norman, D. A. (1988). *The psychology of everyday things* (pp. xi, 257). Basic Books.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., & Musen, M. A. (2009). BioPortal: Ontologies and integrated

data resources at the click of a mouse. *Nucleic Acids Research*, *37*(suppl_2), W170–W173. https://doi.org/10/dm869h

Office of the National Coordinator. (2019). *Health Information Exchange.* HealthIT.gov. https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/health-information-exchange

Ong, E., & He, Y. (2016). Community-based ontology development, annotation and discussion with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia. In *AMIA Summits on Translational Science Proceedings*, *2016*, 65–74.

OpenClinical. (2005). *OpenClinical: Medical terminologies, vocabularies, nomenclatures, coding and classification systems*. Openclinical.Org. http://www.openclinical.org/medicalterminologies.html

Peffers, K., Rothenberger, M., & Kuechler, B. (Eds.). (2012). *Design science research in information systems. Advances in Theory and Practice* (Vol. 7286). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29863-9

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10/cxnmc8

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv:2010.16061 [Cs, Stat]*. http://arxiv.org/abs/2010.16061

Randles, A., Junior, A. C., & O'Sullivan, D. (2021). A vocabulary for describing mapping quality assessment, refinement and validation. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 425–430. https://doi.org/10.1109/ICSC50631.2021.00076

Reason, J. (2000). Human error: Models and management. *British Medical Journal*, *320*(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768

Rogers, E. M., & Marshall, L. R. (2003). *Diffusion of Innovations, 5th Edition*. Free Press.

Rose, J., Fisch, B., Hogan, W., Levy, B., Marshal, P., Thomas, D., & Kirkley, D. (2001). Common medical terminology comes of age, Part One: Standard language improves healthcare quality. *Journal of Healthcare Information Management : JHIM*, *15*, 307–318.

Saitwal, H., Qing, D., Jones, S., Bernstam, E. V., Chute, C. G., & Johnson, T. R. (2012). Cross-terminology mapping challenges: A demonstration using medication terminological systems. *Journal of Biomedical Informatics*, *45*(4), 613–625. https://doi.org/10/f35vt4

Salvadores, M., Alexander, P. R., Musen, M. A., & Noy, N. F. (2013). BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semantic Web*, *4*(3), 277–284.

Schmidt, T. (2021, April 14). You can now open your KNIME hub spaces for collaboration. *KNIME*. https://www.knime.com/blog/knime-hub-collaboration-space

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., Hunter, C. I., Hurwitz, B. L., Kyrpides, N. C., Meyer, F., Mizrachi, I. K., Sansone, S.-A., Sutton, G., Tighe, S., & Walls, R. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data*, *7*(1), 188. https://doi.org/10/gg3knq

Shah, V., Shah, B., Goswami, R., Kumar, S., & Moradiya, C. (2018). Creation of unambiguous centralized knowledge base from UMLS metathesaurus. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1270–1276. https://doi.org/10.1109/BIBM.2018.8621384

Shapiro, J. S., Mostashari, F., Hripcsak, G., Soulakis, N., & Kuperman, G. (2011). Using health information exchange to improve public health. *American Journal of Public Health*, *101*(4), 616–623. https://doi.org/10/d7j6p4

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, *18*(3). https://doi.org/10.13053/cys-18-3-2043

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.

Simperl, B., Paslaru, E., & Mochol, M. (2006). Cost estimation for ontology development.

Business Information Systems. In 9th International Conference on Business Information Systems.

Simperl, E., Bürger, T., Hangl, S., Wörgl, S., & Popov, I. (2012). ONTOCOM: A reliable cost estimation method for ontology development projects. *Journal of Web Semantics*, *16*, 1–16. https://doi.org/10/gnjfjk

Sittig, D. F., & Singh, H. (2010). A new socio-technical model for studying health information technology in complex adaptive healthcare systems. *Quality & Safety in Health Care*, *19*(Suppl 3), i68–i74. https://doi.org/10/cb3j3z

Sittig, D. F., & Singh, H. (2015). A new socio-technical model for studying health information technology in complex adaptive healthcare systems. In V. L. Patel, T. G. Kannampallil, & D. R. Kaufman (Eds.), *Cognitive Informatics for Biomedicine: Human Computer Interaction in Healthcare* (pp. 59–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-17272-9_4

Sittig, D. F., Wright, A., Coiera, E., Magrabi, F., Ratwani, R., Bates, D. W., & Singh, H. (2020). Current challenges in health information technology–related patient safety. *Health Informatics Journal*, *26*(1), 181–189. https://doi.org/10/gh26q9

Smithwick, J. (2015). Unlocking The value of unstructured patient data. *Clinical Leader*. https://www.clinicalleader.com/doc/unlocking-the-value-of-unstructured-patient-data-0002

Sonnenberg, C., & vom Brocke, J. (2012). Evaluations in the science of the artificial – reconsidering the build-evaluate pattern in design science research. In K. Peffers, M. Rothenberger, & B. Kuechler (Eds.), *Design Science Research in Information Systems. Advances in Theory and Practice* (pp. 381–397). Springer. https://doi.org/10/gj62m8

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., & Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. *Bioinformatics*, *34*(11), 1962–1965. https://doi.org/10/gdk4vz

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10/gf38jv

Tichy, W. F. (1998). Should computer scientists experiment more? *Computer*, *31*(5), 32–40. https://doi.org/10.1109/2.675631

Ure, J., Procter, R., & Lin, Y. (2008). *A socio-technical perspective on ontology development in healthgrids*. In *Proceedings of UK eScience All Hands Meeting* (10-13)

Vaishnavi, V. K., Kuechler, W., & Kuechler, W. (2015). *Design science research methods and patterns: Innovating information and communication technology* (2nd ed.). Taylor & Francis Group.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. JSTOR. https://doi.org/10/gc8zn2

Wang, P., Hu, Y., Bai, S., & Zou, S. (2021). Matching biomedical ontologies: Construction of matching clues and systematic evaluation of different combinations of matchers. *JMIR Medical Informatics*, *9*(8), e28212. https://doi.org/10.2196/28212

Wieringa, R. (2009). Design science as nested problem solving. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 1. https://doi.org/10/bn3vzv

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10/bdd4

Yan, C., Zhang, Y., Liu, K., Zhao, J., Shi, Y., & Liu, S. (2021). Enhancing unsupervised medical entity linking with multi-instance learning. *BMC Medical Informatics and Decision Making*, *21*(9), 317. https://doi.org/10.1186/s12911-021-01654-z

Zeng, M. L. (2008). Knowledge organization systems (KOS). *Knowledge Organization*, *35*(2), 160–182. https://doi.org/10/gf8692

Zeng, M. L. (2019). Interoperability. *Knowledge Organization*, *46*(2), 122–146. https://doi.org/10/gf3b24

Zeng, M. L., & Clunis, J. (2020). FAIR + FIT: Guiding principles and functional metrics for linked open data (LOD) KOS products. *Journal of Data and Information Science*, *5*(1). https://doi.org/10.2478/jdis-2020-0008

Zeng, M. L., Hong, Y., Clunis, J., He, S., & Coladangelo, L. P. (2020). Implications of knowledge organization systems for health information exchange and communication during the COVID-19 pandemic. *Data and Information Management*, *4*(3), 148-170. https://doi.org/10.2478/dim-2020-0009

Zhang, S., & Bodenreider, O. (2007). Experience in aligning anatomical ontologies. *International Journal on Semantic Web and Information Systems*, *3*(2), 1–26.

Zhou, L., Plasek, J. M., Mahoney, L. M., Chang, F. Y., DiMaggio, D., & Rocha, R. A. (2012). Mapping partners master drug dictionary to RxNorm using an NLP-based approach. *Journal of Biomedical Informatics*, *45*(4), 626–633. https://doi.org/10/dzksz3

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., & Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, *46*(2), 200–211. https://doi.org/10/f4tpr2