THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Data Linkage for Querying Heterogeneous Databases

Mohammed Abdul Salam, GOLLAPALLI

Master of Information Technology (MIT) & Bachelor of Information Technology (BIT)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2013

School of Information Technology & Electrical Engineering

Abstract

Data Linkage is an important step that can provide valuable insights for evidence-based decision making, especially for crucial events. Performing sensible queries across heterogeneous databases containing millions of records is a complex task that requires a complete understanding of each contributing database's schema to define the structure of its information. The key aim is to approximate the structure and content of the induced data into a concise synopsis in order to extract and link meaningful facts. Current techniques primarily focus on performing pair-wise attribute matching and pay little attention in discovering direct and weighted cluster correlations for linking semantic equivalent datasets. We identify such problems as four major research issues in Data Linkage: associated costs in pair-wise matching, record matching overheads, semantic flow of information restrictions, and single order classification limitations.

In this doctorial dissertation, we introduce a new multi-faceted classification technique for performing structural analysis on knowledge domain clusters, using a novel Ontology Guided Data Linkage (OGDL) framework. In order to support self-organization of contributing databases through the discovery of structural dependencies, we introduce a series of algorithms for performing multi-level exploitation of ontological domain knowledge relating to tables, attributes and tuples. These techniques are of great help for automating the discovery of schema structures across multiple databases, based on the use of direct and weighted correlations between different ontological concepts, using a novel h-gram (hash gram) record matching technique for concept clustering and cluster mapping. Moreover, through a set of accuracy, performance and scalability experimental tests run on real-world datasets, we demonstrate the feasibility of our OGDL algorithms and show that our framework runs in polynomial time and performs well in practice.

Data Linkage is an important enabling technology in eHealth as linked data is a cost effective approach towards advancing research outcomes into health policies, detect any adverse drug reactions, reduce costs, and uncover any non-practices within the health system. Hence, to illustrate the efficiency and effectiveness of OGDL in real-world applications, we comprehensively used clinical risk management domain as our practical example. For this reason, we further extended our OGDL framework and introduced a composite clinical risk management success indicator data linkage, which consists of clinical risk factors combined with clinical resource and intervention factors that have shown to be associated with good and safe patient outcomes and with quality health care. The aim is to introduce a novel primitive upper ontology for semantic interoperability of health data and subsequent clinical risk management, and use it to map patient case data to reason about problems and solutions. Our experiments are performed on the Australian emergency medicine clinical trial datasets, demonstrating an effective method for the creation of a new risk management approach using semantic interoperability and reasoning.

The main contributions of this thesis include: introducing a novel h-gram record matching technique highly reducing the number of comparisons required in determining entity similarities, providing a highly effective and efficient OGDL framework for querying and integrating heterogeneous databases in the presence of data uncertainties, demonstrating an effective method for identifying how different sets of tables, attributes and tuples can be linked with the primary aim to understand the past and predict the future, providing a method for discovering ontological instances in domain specific clusters that reveals how different sets of information is organized to support information flow, introducing a novel primitive upper ontology for semantic interoperability, and finally supporting the development of a best-practice clinical practice guideline assessment framework with evidence based on the collaboration platform's health knowledge repository.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the *Copyright Act 1968*.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

1. M. Gollapalli, X. Li, I. Wood, and G. Governatori, Ontology Guided Data Linkage for Discovering Meaningful Data Facts, Springer, Advanced Data Mining and Applications (ADMA), Volume No: LNCS 7121, Beijing, Dec 2011, pp. 252-265.
2. M. Gollapalli, X. Li, and I. Wood, Automated discovery of multi-faceted ontologies for accurate query answering and future semantic reasoning, Elsevier, J. of Data and Knowledge Engineering (DKE), May 2013, pp. DATAK-01441.
3. M. Gollapalli, X. Li, I. Wood, and G. Governatori, Approximate Record Matching Using Hash Grams, IEEE, International Conference on Data Mining (ICDM) Conference Workshop, Vancouver, Dec 2011, pp. 504-511.
4. M. Gollapalli, and X. Li, A Framework of Ontology Guided Data Linkage for Evidence based Knowledge Extraction and Information Sharing, 29th IEEE International Conference on Data Engineering (ICDE) PhD Symposium, Brisbane, Apr 2013.
5. L. New, M. Gollapalli, and X. Li, A Data Driven Approach to Holistic Dynamic Clinical Risk Indicator Development for Best-Practice Evidence-Based Decision Support at the Point of Care, (on hold for IP)
6. M. Gollapalli, L. New, and X. Li, An Efficient Clinical Upper Ontology Cloud Framework for Collaborative Best-Practice Success Indicator Development and Dissemination, (on hold for IP)
7. M. Gollapalli, and X. Li, Classification of Approximate Data Linkage Techniques, World Scientific, Int. J. of Semantic Computing (IJSC), submitted Mar 2012 (under peer review)

Publications included in this thesis

Contributor	Statement of contribution
M. Gollapalli, X. Li, I. Wood, and G. Governatori, Ontology Guided Data Linkage for Discovering Meaningful Data Facts, Springer, Advanced Data Mining and Applications (ADMA), Volume No: LNCS 7121, Beijing, Dec 2011, pp. 252-265. - Incorporated as sub-sections in Chapter 1 & 4	
Mohammed Gollapalli	Wrote the paper (90%) Designed & conducted experiments (90%)
Xue Li	Reviewed & edited the paper (45%)
Ian Wood	Reviewed & edited the paper (20%)
Guido Governatori	Reviewed the paper (5%)
M. Gollapalli, and X. Li, Classification of Approximate Data Linkage Techniques, World Scientific, Int. Journal of Semantic Computing (IJSC), submitted Mar 2012 (under review) - Incorporated as Chapter 2	
Mohammed Gollapalli	Wrote the paper (90%)
Xue Li	Reviewed & edited the paper (25%)
M. Gollapalli, X. Li, I. Wood, and G. Governatori, Approximate Record Matching Using Hash Grams, IEEE, International Conference on Data Mining (ICDM) Conference Workshop, Vancouver, Dec 2011, pp. 504-511. - incorporated as Chapter 3	
Mohammed Gollapalli	Wrote the paper (90%) Designed & conducted experiments (90%)
Xue Li	Reviewed & edited the paper (25%)
Ian Wood	Reviewed & edited the paper (10%) Statistical analysis of data in Tables 3.3 (45%)
Guido Governatori	Reviewed the paper (5%)
M. Gollapalli, X. Li, and I. Wood, Automated discovery of multi-faceted ontologies for accurate query answering and future semantic reasoning, Elsevier, J. of Data and Knowledge Engineering (DKE), May 2013, pp. DATAK-01441.	

<table>
<tr><td colspan="2">- Incorporated as sub-sections in Chapter 1 and 4</td></tr>
<tr><td>Mohammed Gollapalli</td><td>Wrote the paper (95%)
Designed & conducted experiments (90%)</td></tr>
<tr><td>Xue Li</td><td>Reviewed & edited the paper (25%)</td></tr>
<tr><td>Ian Wood</td><td>Statistical analysis of data in Tables 3-3, 5-1 (35%)</td></tr>
<tr><td colspan="2">L. New, M. Gollapalli, and X. Li, A Data Driven Approach to Holistic Dynamic Clinical Risk Indicator Development for Best-Practice Evidence-Based Decision Support at the Point of Care, (on hold for IP).
- Incorporated as sub-sections in Chapter 1 &amp; 5</td></tr>
<tr><td>Mohammed Gollapalli</td><td>Wrote &amp; edited the paper (25%)
Designed & conducted experiments (80%)</td></tr>
<tr><td>Lisa New</td><td>Wrote &amp; edited the paper (75%)
Conducted experiments (20%)</td></tr>
<tr><td>Xue Li</td><td>Reviewed the paper (20%)</td></tr>
<tr><td colspan="2">M. Gollapalli, L. New, and X. Li, An Efficient Clinical Upper Ontology Cloud Framework for Collaborative Best-Practice Success Indicator Development and Dissemination, (on hold for IP).
- Incorporated as sub-sections in Chapter 1 &amp; 6</td></tr>
<tr><td>Mohammed Gollapalli</td><td>Wrote &amp; edited the paper (25%)
Designed & conducted experiments (80%)</td></tr>
<tr><td>Lisa New</td><td>Wrote &amp; edited the paper (75%)</td></tr>
<tr><td>Xue Li</td><td>Reviewed the paper (20%)</td></tr>
<tr><td colspan="2">M. Gollapalli, and X. Li, A Framework of Ontology Guided Data Linkage for Evidence based Knowledge Extraction and Information Sharing, 29th IEEE International Conference on Data Engineering (ICDE) PhD Symposium, Brisbane, Apr 2013.
- Incorporated as sub-sections in Chapter 1, 4 &amp; 6</td></tr>
<tr><td>Mohammed Gollapalli</td><td>Wrote the paper (95%)
Designed & conducted experiments (90%)</td></tr>
<tr><td>Xue Li</td><td>Reviewed the paper (50%)</td></tr>
</table>

Contributions by others to the thesis

I would like to thank Dr. Lisa New, with whom I had numerous discussions in the final stages of my research progress. Dr. New helped me in applying my research work into clinical risk management domain. She helped me in extending the OGDL Framework with her First-Order Logic Primitive Upper Ontology for Risk Management (FLORM) Formal Language, for the purpose of collaborative development and user-friendly querying of a shared risk management knowledge repository (CBOK) in a semantic web expert system application for real-time risk management decision-support (SWARM). Her SWARM Cloud Framework will be realised through OGDL extensions to include novel semantic web technology RDF triplet collaboration threads of real-time and historic communication of opinions and facts, viewable in filterable interactive argument trees to support transparent real-time discourse ethics by experts. This includes capacity to support approximated expert consensus formulation on best-practice risk management and related resource application and coordination, within data mined windows-of-opportunities for risk prevention and mitigation, given historic patterns of complex problems and relevant holistic solutions.

Statement of parts of the thesis submitted to qualify for the award of another degree

No Submissions made for any other degree.

Acknowledgements

First and foremost, I take immense pleasure in thanking my supervisor, Dr. Xue Li for the valuable guidance, support and advice he has provided throughout the course of my research. Dr. Li inspired me greatly to work in this research and has motivated me tremendously in achieving my milestones. I had the pleasure of sharing my views with Dr. Li whose words, over the years has taught me much about understanding research in a ‘problem-driven’ approach and the ways in handling it. I would also like to gratefully thank my second supervisor, Dr. Ian Wood from the School of Mathematics & Physics for the technical discussions, help with experimental setup and general advice he has provided vital for the success of my research. I also would like to express my gratitude to Dr. Guido Governatori and other staff members, friends, and colleagues who rendered their help during the period of my research work.

No one walks alone on the journey of life. Much of what I have contributed in my research over the years came as the result of being a husband of my beloved wife, Sajeeda and father of 2 wonderful and delightful daughters, Sadiqah and Ayesha, who in their own ways inspired me and, morally encouraged me a tremendous amount to the content of my research. I am forever indebted especially to my wife Sajeeda for her understanding, endless patience and encouragement when it was most required. I would also like to express my heartfelt thanks to my beloved parents for their blessings and wishes for the successful completion of this research.

Finally, I would like to convey my special thanks to the University of Queensland and Faculty of Information Technology & Electrical Engineering (ITEE) for providing the financial means and research facilities.

Keywords

Concept Modeling, Data & Knowledge Visualisation, Data Linkage, Decision-Support, eHealth, Query Processing, Risk Management, Semantic Interoperability, Semantic Reasoning, Upper Ontology, Cloud Computing

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC 080109 – 40% - Pattern Recognition and Data Mining
ANZSRC 080403 – 35% - Data Structures
ANZSRC 080605 – 25% - Decision Support and Group Support Systems

Fields of Research (FoR) Classification

FoR 0801 – 30% - Artificial Intelligence and Image Processing
FoR 0804 – 35% - Data Format
FoR 0806 – 35% - Information Systems

Table of Contents

List of Figures

List of Tables

List of Abbreviations

5WH	Who, What, Where, When, Why, How
ARFF	Active Relation File Format
BF	Brute Force approach
BI	Business Intelligence
CBOK	Collective Body of Knowledge
CPG	Clinical Practice Guidelines
DL	Data Linkage
DSS	Decision Support System
ED	Emergency Department
EHMP	Ecosystem Health Monitoring Program
EKW	Enterprise Knowledge Warehouse
FLORM	First Order Logic-Primitive Upper Ontology for Risk Management
GS	Gold Standard
HL7	Health Level Seven International
h-gram	Hash Gram Record Matching Technique
ICP	Integrated Care Pathways
MS	Microsoft
OGDL	Ontology Guided Data Linkage Framework
QBE	Query by Example
RDF	Resource Description Framework
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
SQL	Structured Query Language
SWARM	Semantic Web Expert System Application for Risk Management
TF	Term Frequency
TWB	The World Bank
WHO	World Health Organisation
WWF	World Wildlife Fund

إِنَّا خَلَقْنَا ٱلْإِنسَٰنَ مِن نُّطْفَةٍ أَمْشَاجٍ نَّبْتَلِيهِ فَجَعَلْنَٰهُ سَمِيعًۢا
بَصِيرًا ﴿٢﴾
إِنَّا هَدَيْنَٰهُ ٱلسَّبِيلَ إِمَّا شَاكِرًا وَإِمَّا كَفُورًا ﴿٣﴾

We created Man from a drop of mingled sperm, in order to try him: So We gave him (the gifts), of Hearing and Sight. We showed him the Way: whether he be grateful or ungrateful (rests on his will).

(Holy Quran, Chapter 76 (The Man): Verses 2-3)

Chapter 1
Introduction

1.1. Research Problem

Organizations worldwide have been collecting data for decades. The World Bank [24], The National Climatic Data Centre [49], and countless other private and public organizations have been collecting, storing, processing and analysing massive amounts of data which has the potential to be linked for the discovery of underlying factors to critical problems. Sharing of large databases between organisations is also of growing importance in many data mining projects, as data from various sources often has to be linked and aggregated in order to improve data quality, or to enrich existing data with additional information [7]. When integrating data from different sources to implement a data warehouse, organisations become aware of potential systematic differences, limitations, restrictions or conflicts which fall under the umbrella-term data heterogeneity [34]. Poor quality data has also been prevalent in databases due to a variety of reasons, including typographical errors, lack of standards etc. To be able to query and integrate data in the presence of such data uncertainties as depicted in Figure 1.1, a central problem is the ability to identify whether heterogeneous database tables, attributes and tuples can be linked with the primary aim to understand the past and predict the future.

In response to the aforementioned challenges, significant advances have been made in recent years in mining structures of databases with the aim to acquire crucial fact finding information that is not otherwise available, or that would require time-consuming and expensive manual procedures. Schemas are definitions that identify the structure of induced data and are the result of a database design segments. The relational database schemas that are invariant in time hold valuable information in their tables, attributes and tuples which can aid in identifying semantically similar objects. The process of identifying these schema structures has been one of the essential elements of data mining process [21-26]. Accurate integration of heterogeneous database schema can provide valuable insights that are useful for evidence-based decision making, especially for crucial events. In the schema integration process, each individual database can be analysed to provide and extract local schema definitions of the data. These local schema definitions can be used for the development of a global schema which integrates and subsumes the local schema in such a way that (global) users are provided with a uniform and correct view of the global database [19]. With the help of global schema structures, we can derive hierarchical relationships up to the instance level across datasets. However, without having this global schema, extracting meaningful data into a usable form can become a tedious process [5, 8, 14, 18, 21, and 26]. Traditional local-to-global schema-based techniques lack the ability to allow computational linkage and are not suitable when dealing with heterogeneous databases [2, 5, 8, 18, 57, 61 and 66]. To make things worse, the data could be “dirty” and differences might exist in the structure and semantics maintained across different databases. Research communities have also stressed Schema Pattern Matching [21 to 26] and SQL Querying [27, 28]. Schema Pattern Matching uses database schema to devise clues as to the semantic meaning of the data. Constraints are used to define requirements, generated by hand or through a variety of tools. However, the main problems with Schema Pattern Matching are insufficiency and redundancy.

Data linkage (also known as data matching, probabilistic matching, and instance identification) is the process of identifying records which represent the same real world entity despite typographical and formatting constraints [18, 25, 32, 34, and 37]. In conducting our research, we observed four prime areas where data linkage is a persistent, yet heavily researched problem:

- Medical science for DNA sequence matching and biological sequence alignment [12, 18, 21, 47, 56, and 80-84];
- Government departments for taxation and payout tracking [5, 24, 30, 48, and 79];
- Businesses integrating the data of acquired companies into their centralized systems [2, 36, and 42];
- Law enforcement for data matching across domains, such as banking and the electoral commission [24, 30, 33, 49, and 50].

Traditional data linkage approaches use similarity scores that compare tuple values from different attributes, and declare it as matches if the score is above a certain threshold [2, 10, 18, 61, 67, and 79]. These approaches perform quite well when comparing similar databases with clean data. However, when dealing with a large amount of variable data, comparison of tuple values alone is not enough [1, 2]. It is necessary to apply domain knowledge when attempting to perform data linkage where there are inconsistencies in the data. The same problem applies to database migrations, and to other data intensive tasks that involve disparate databases without common schemas. Furthermore, the creation of data linkage between heterogeneous databases requires the discovery of all possible primary and foreign key relationships that may exist between different attribute pairs, on a global spectrum [1, 3, 8, 11, and14-16].

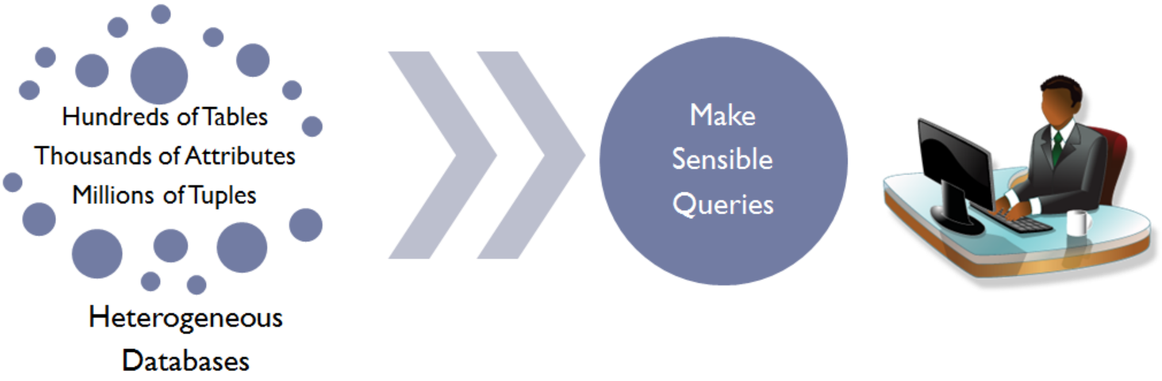


Figure 1.1: Use of Heterogeneous Databases for Data Linkage

1.1.1. Role of Ontologies in Data Linkage

One of the most common problems in discovering global database schemas is semantic heterogeneity-if it is not detected and resolved, the usage of integrated data leads to invalid results [19]. Furthermore, the invalid results could become undetected especially when dealing with large quantities of heterogeneous databases. An Ontology typically provides a vocabulary describing a domain of interest and a specification of the meaning of terms in that vocabulary [2]. Ontology is increasingly seen as a key factor for enabling interoperability across heterogeneous systems and semantic web applications [32]. Ontologies are expected to play a significant role in various application domains in the emerging Semantic Web, linking databases semantically. Furthermore, the ability to efficiently and effectively perform ontology reuse is commonly acknowledged to play a crucial role in the large-scale dissemination of ontologies and ontology-driven technologies [6]. Eusenat et al in [2] has shown that such ontology based methods can be highly effective, when combined with other methodologies.

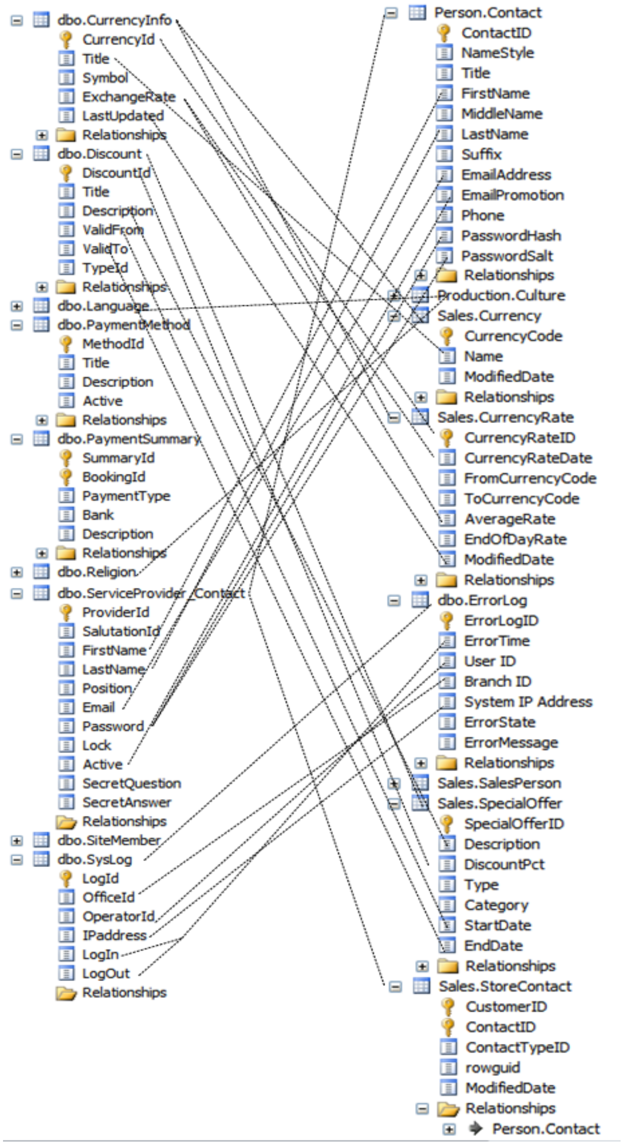


Figure 1.2: The comparison of database schemas with similar ontologies

A key step in the integration of databases is the identification of semantic correspondences among ontology attribute pairs [2, 4, and 5]. Therefore, this research will primarily focus on the identification of semantic data coordination, using ontology matching principles. However, ontology matching at an attribute level can be very expensive and have varying relevance. For instance, a table or an attribute can have multiple ontologies, as shown in Figure 1.2, which demonstrates ontology correspondences as references between two input schema's table attributes. As can be seen from the diagram, it depicts two input schemas with similar ontologies: on the left there is a representation of an 'online transaction processing' database with data of the provision of discounts and special offers, in multiple countries and in multiple currencies. On the right there is a representation of its 'data warehouse equivalent', used to develop various business intelligence (BI) reports and to perform data modelling, as well as a number of other data mining tasks. The dotted arrows in Figure 1.2 indicate tables and attributes matching instances between multiple schemas and multiple ontologies. For instance, the ‘title’ attribute from the ‘dbo.CurrencyInfo’ table is referenced to the ‘name’ attribute in the ‘sales.Currency’ table. Figure 1.2 further illustrates that schemas overlap each other in general, and that each schema can also have unique information, not present in any other schema (for example, ‘currencies’ and ‘exchange rates’).

1.1.2. Role of Probabilistic Techniques in Data Linkage

Exact Matching techniques can give more insight into the content and meaning of schema elements [25, 31]. Exact matching uses a unique identifier present in databases being compared. The unique identifier can only be linked to one individual item, or an event (for example, a driver’s license number). The Exact Matching technique is helpful in situations where the data linkage to be performed belongs to one data source. However, exact matching comparison does not suffice for matching records when the data contains errors, for example typographical mistakes, or when the data have multiple representations, such as through the use of abbreviations or synonyms [10]. Unlike Exact matching, Probabilistic techniques [34-36] are used to perform Data Linkage on a likelihood basis (i.e. performing matching based on the success threshold ratio). Output results can vary in different formats such as “match, possible match, and non-match” basis, Boolean type true or false match basis, nearest and outermost distance match basis, discrete or continuous match basis etc.

In summary, we consider the problem of discovering ontological instances in domain specific clusters that reveals how different tables, attributes and tuples are organized between and within databases to support information flow. Inspired by this, in this thesis, we consider a new type of data linkage approach, namely, the exploitation of hidden relationships between tables, attributes and tuples towards knowledge discovery at different levels of data abstraction, including the ontology, schema and instance levels. Essentially, we introduce a new approach of linking data across heterogeneous databases based on approximate data matching to identify correspondences between related entities which assist in discovering meaningful data relationships. The work presented in this thesis belongs to the research domain of mining heterogeneous data. The study is motivated by the application of clinical risk indicator analysis. However, generalised research problems can also serve the studies in other application domains.

1.2. Motivation and Scope

According to our observation, the most intuitive way for linking heterogeneous databases is to develop a schema structure abstract providing a more logical information flow view. While this approach seems plausible, it may nevertheless still fail if: 1) pair-wise matching restrictions exist; 2) not enough data quality checks are made; 3) if the user is unable to identify any errors; or 4) if the semantic flow of linked data is not correct through automated process.

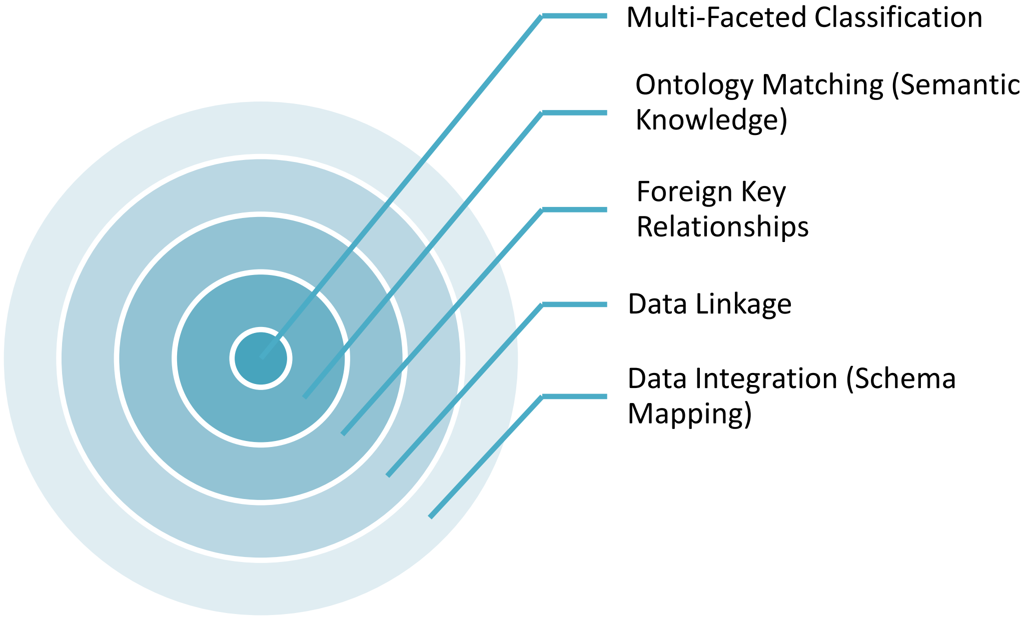

Multi-Faceted Classification
Ontology Matching (Semantic Knowledge)
Foreign Key Relationships
Data Linkage
Data Integration (Schema Mapping)

Figure 1.3: OGDL Research Gap Visualisation

We identify these problems as four major research issues in Data Linkage as depicted in Figure 1.3: associated costs in pair-wise matching, Semantic flow of information restrictions, Record matching overheads and Single Order Classification.

Pair-wise matching costs: The fundamental problem that arises each time in performing data linkage on large volumes of heterogeneous databases is to discover all possible relationships based on matching similar tuple values that might exist between each attribute pairs. Pair-wise matching of attributes [34, 35, and 36] between different data source tables is a suitable approach for small databases. However, real-world data collected from enterprise organizations can have hundreds of tables and thousands of columns. Hence, performing pair-wise attribute matching can be highly expensive in terms of associated computational costs, which is perhaps the main drawback found in existing data linkage methods, and this is also what restricts its performance in terms of accuracy. In order to reduce the number of pair-wise comparisons, we employ a ‘multi-layer’ ontology-based clustering technique, by modelling large amounts of input information into high-density clusters at different levels. As some chains-of-relationships have stronger correlation weights than others, we focused our research on the identification of such correspondences between crucial attributes, together with its semantic information flow.

Semantic flow anomalies: Semantic information is used as data abstraction principles to perform data linkage. The development of a novel system that embodies this approach faces a number of challenges. In this thesis, our solution to handle these challenges will integrate a variety of approaches, by extending existing methods and proposing new multi-faceted strategy ontology guided data linkage (OGDL) framework. The proposed framework uses different datasets as input and performs data uncertainty analysis for data cleaning and to organise data into homogeneous strata groups. The strata samples are used to form different cluster levels. The framework then performs cluster stem-and-leaf joins, using a multi-faceted cluster mapping technique. These results are further analysed to construct hierarchical cluster mapping trees. The ontological structures are summarized as candidate, primary, partial, and foreign key relational data (linkage) relationships. The final results are further integrated into knowledge based data analysis tools to support sensible query making to discover meaningful and accurate data facts.

Record matching overheads: Small inconsistencies in records can prevent matching between two otherwise identical set of records. To deal with this problem, probabilistic approaches are often performed on a likelihood basis (i.e. performing matching based on the success threshold ratio). In this thesis, we present a novel h-gram (hash gram) [79] technique for probabilistic record matching. The h-gram technique is aimed at reducing the runtime costs when comparing records, and to get probabilistic results in a timely manner. The h-gram matching process extends to traditional n-grams by the transformation of the grams into equivalent numerical realities, thus overcoming the disadvantages of random-assignation hashing systems. It also provides more options for gram scaling and for error threshold tolerance. This is similar to the approach taken by [8], although we do not store hash codes of all the sample data. We reduce the cost associated with record matching by utilizing scale based hashing; increasing matching probability through fine turning; and by reducing the cost associated with the storage of most frequent hash codes of matching records. We employed the h-gram technique within our OGDL framework to create and correlate clusters at different levels and thus significantly improving the OGDL framework's performance.

Unfit single order modelling: Our research findings suggest that single order classification of data does not provide the necessary flexibility to accurately define semantic mappings of variables. For instance, different organizations typically maintain different rules and standards for storage of their business data, and there are instances of such databases being poorly designed, and/or without data models. Platform independent databases that target the global marketplace have also emerged in recent years. The variability of the quality of such data sources leads to the risk that the semantic flow of the data (as per their relationships) is not in a fixed direction. In order to increase the probability of discovering correlated clusters, in this thesis, we applied a 'multi-faceted ontology-based cluster mapping' strategy. The overarching objective is to develop the ontological domain information as represented in its tables, attributes and tuples, in multiple facets (arrangements), instead of by a predetermined order. The aim is also to capture the flow of meaningful semantic data and to concurrently construct self-expanding hierarchical semantic tree structures, which is crucial for high quality data linkage.

In this thesis, we have described methods for constructing three different kinds of representations: sequential; parallel; and mixed facets. A sequential facet aims to classify data based on the ontological findings of table level clusters, followed by attribute level clusters and then tuple level clusters. A parallel facet does not prioritize any sequence order, and equally classifies data based on the chance of finding pairs within table level clusters; within attribute level clusters or within tuple level clusters. A mixed facet classifies data through combined cross referencing at the table, attribute and tuple cluster levels. The obtained results are further narrowed down in order to discover candidate keys, primary keys, foreign keys, and partially related keys. These results can be integrated with IBM or Microsoft's Query-by-Example (QBE) tools with the aim to make sensible queries that discover meaningful and accurate (data) facts.

1.3. Research Background

The problem of extracting semantic structures from heterogeneous databases can be addressed at different levels of complexity. Pure semantic based extraction, using thesauri based dictionaries, presents one extreme [2, 9]. Problem formulation based on syntactic approaches presents the other extreme. In general, many sophisticated data linkage techniques have been applied which can be broadly classified into deterministic, probabilistic and modern approaches [2]. In the past, iterative techniques have constantly been proposed, such as 'Iterative Deduplication' [10], 'Parallel Linkage' [11], and 'DNA Sequence' [12]. Findings suggest that they may produce accurate results, but that it comes with an additional cost. The advantages of these techniques include decreasing false positive rates, but can be expensive computationally; the similarity comparison is not limited to attribute comparisons only; and it has to cope with continuously updated distance metrics, as each new duplicate correlation is discovered.

Michel Gagnon [13] proposed a local to global ontology mapping method for integrating data sources. While this technique helps to achieve a global ontology picture, it does not make use of schema mapping strengths and does not have the capacity to understand tuples. The authors in [14] and [15] have proposed global schema mapping as a resolution for data linkage. However, schema mapping by itself is not sufficient [2] and is not a panacea to the identification of semantic structures of unrelated databases.

CORDS [16] constitutes a substantial contribution to quality query-based approaches. CORDS discovers correlations and soft functional dependencies through an automated process, using column pairs. CORDS creates column groups through a series of processes that include enumeration of candidate pairs; elimination of unlikely candidates; and statistical analysis to identify correlations. Unfortunately, CORDS primarily uses relational database architecture and it performs pair-wise attribute matching instead of pursuing a structural level approach. Pair-wise attribute matching is highly expensive when applied to large volumes of unrelated databases, and does not discover semantic mappings.

iDisc [5] creates database representations using a multi-process learning technique. Base clusters are used to uncover topical clusters, which are then aggregated through meta-clustering. The advantage of the iDisc framework is that it supports the extension of existing clusters and representations. However, the iDisc approach doesn't support reasoning based mapping (which in itself could be described as a cumbersome approach).The iDisc approach also doesn't consider the building of ontological structure mapping trees.

In this research, we introduce a multi-faceted classification technique for performing structural analysis on knowledge domain clusters, using a novel Ontology Guided Data Linkage (OGDL) framework [1]. The framework supports self-organization of contributing databases through the discovery of structural dependencies, by performing multi-level exploitation of ontological domain knowledge relating to tables, attributes and tuples. The framework thus automates the discovery of schema structures across unrelated databases, based on the use of direct and weighted correlations between different ontological concepts, using our proposed h-gram (hash gram) record matching technique for concept clustering and cluster mapping. We demonstrate the feasibility of our OGDL prototype and algorithms through a set of accuracy, performance and scalability experimental tests run on real-world data, and show that our system runs in polynomial time and performs well in practice.

This research will then introduce OGDL framework's advancement towards clinical risk management success indicator development, which consists of clinical risk factors combined with clinical resource and intervention factors that have shown to be associated with good and safe patient outcomes and with quality health care. The aim is to introduce a novel primitive upper ontology for semantic interoperability of health data and subsequent clinical risk management, and use it to map patient case data to reason about problems and solutions. Our experiments are performed on the Australian emergency medicine clinical trial datasets, and demonstrate the creation of a new risk management approach using semantic interoperability and reasoning. This work has applicability to eHealth applications for dynamic clinical decision-support; and for equitable health service planning, funding and delivery. This research is significant for stakeholders of health improvement and health service provision. It also has wider applicability of semantic web based collaborative risk management with real-world real-time dynamic data flow, supported by artificial intelligence.

1.4. Research Contributions

The main contributions of this thesis include:

1. The first contribution of this research addresses our approach towards data uncertainty and data cleaning. In order to perform a successful data linkage between disparate noisy datasets, the data needs to be organized in a format that supports user-friendly access to different sets and subsets of data. Prior to the data linkage process, the proposed data uncertainty process organizes variable datasets into a uniform representation. We introduce a novel Multi-Modular Neural Networks, using the process of ranking and classifying ontological characteristics in multiple modules.

2. The second contribution of this research is our newly proposed h-gram (hash gram) record matching technique [79]. The h-gram record matching is highly significant and advances set-of-sets technique [8] by extending the features of scale based hashing and n-gram techniques. The peculiarity of h-gram matching is that it allows for multiple level of detailed analysis and is not limited to any range of data type or size. h-gram matching is an interactive technique supporting users to build their own data linkage models by tuning system parameters.

3. The third contribution is the introduction of our OGDL framework. The OGDL framework creates multi-layer clusters within each sample set, based on its ontological essence. The clusters self-expand through the application of a multi-faceted cluster mapping strategy, applied on a global spectrum. The framework results are further drilled-down to create schema structures. The resulting schema structures can easily be integrated in existing data mining tools to enhance knowledge discovery. Our contribution presents an extensive evaluation of the OGDL framework as applied to real-world databases in experimental tests for accuracy, performance and scalability analysis.

4. The fourth contribution is the extension of OGDL framework [1] through a novel First-Order Logic Primitive (with less than 100 elements) Upper Ontology for Risk Management, (FLORM), to support development of a risk knowledge repository; and to enable semantic reasoning to deliberate consensus on improved success

and risk indicators. We extend OGDL to extract semantic cluster patterns of past evidence of resource and intervention success for specific problems from the knowledge repository, which is organized through FLORM in a problem-solution framework. This enables machine learning of data driven composite holistic success indicators from the knowledge repository, as an integration of risk indicators with successful resource and intervention indicators. This is significant for an evidence-based approach to risk management.

Our research can benefit data managers, researchers, or analysts, for a variety of purposes, including optimized multi-domain knowledge representation, as it doesn't require a data structure or complex query knowledge. We have addressed high computational overheads through a multi-layer strategy which significantly reduces the amount of data considered for comparison at subsequent stages, and which enables cluster self-expansion through the construction of ontology guided data linkage structures.

Furthermore, our research can be directly applied by the health service planning and management professionals, and health workers, as an introduction to a new approach to collaborative optimised preventive risk management, using a semantic web based collaboration platform for risk management. The approach has wider applicability to public and environmental risk management.

1.5. Dissertation Organization

The thesis organisation is shown in Figure 1.4.

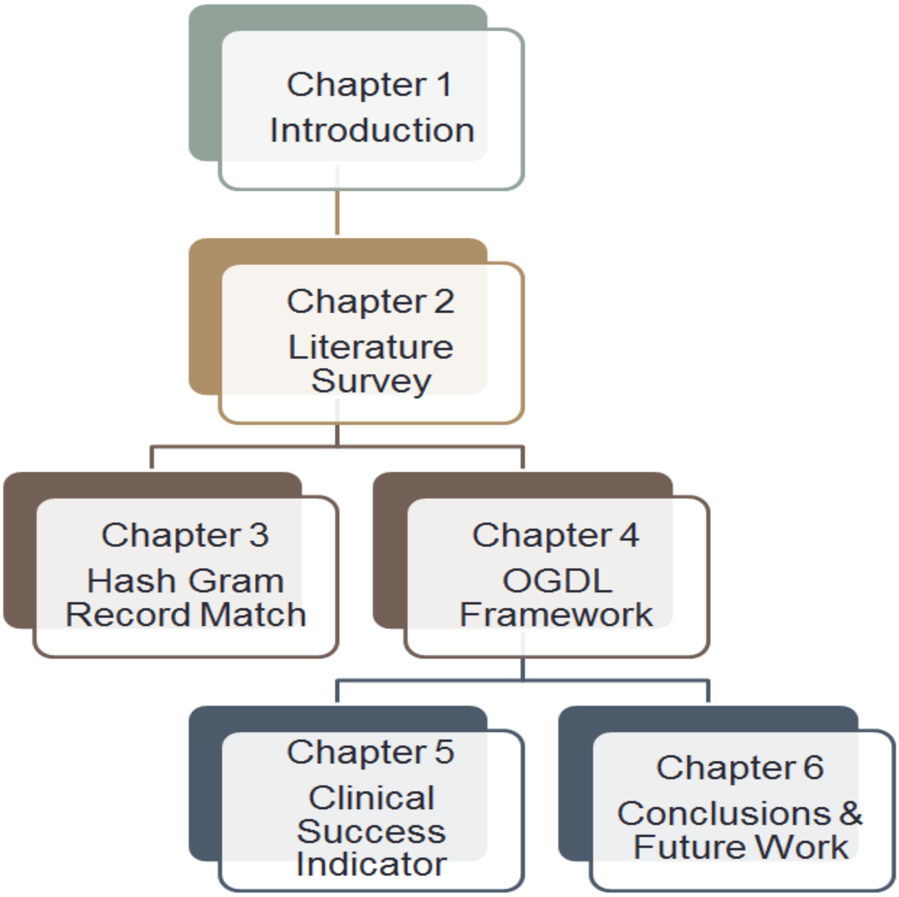


Figure 1.4: Thesis organisation

The introduction chapter introduces Data Linkage and its applications, challenges, and problems in the domain. It also describes the motivations of this research and the methodologies used in this research. Our contributions are summarised in the introduction.

In Chapter 2, we give a literature survey of research in Data Linkage. The purpose for this survey is to establish a basic understanding of Data Linkage, and to discuss the background to our research. Particularly, we focus on the literature related to the work in this thesis, including the topics of SQL based Matching, Exact Matching and Approximate Matching algorithms. Their efficiency, functionality and limitations are critically analysed.

In Chapter 3, a new record matching functionality h-gram (hash gram) technique and its corresponding implementation in Data Linkage is proposed. Intuitive examples and experimental results in terms of the accuracy and performance are also provided.

In Chapter 4, we describe our approach to resolve the research problem by introducing the Ontology Guided Data Linkage (OGDL) Framework. We start by showing the functionality of Data Uncertainty analysis as part of the data preparation stage. For this, a real life classification case (The World Bank [24]) is used to demonstrate how our approach can be utilised and evaluated. We then formally introduce the data linkage problem through our new framework. We discuss the overall experimental design of the experimental technique. We then illustrate performing semantic queries on the obtained results.

In Chapter 5, we introduce OGDL framework’s advancement towards clinical risk management. We provide a data driven approach to holistic dynamic clinical success indicator development for Best-Practice Evidence-Based Decision-Support at the Point of Care. We propose a new approach to derive composite data driven clinical success indicators from a clinical trial dataset, and compare the results with published indicators from existing clinical guidelines. We propose that data driven clinical risk and related resource and implementation indicators be identified through machine learning of past evidence. Our approach can support data integration in a knowledge repository with greatly enhanced data mining capacity, and can enable user-friendly First Order Logic querying to extract meaningful facts without expert IT knowledge and skills.

In Chapter 6, we introduce our conclusions and recommendations for future work in which some potential future research directions are discussed in detail.

1.6. Summary

In this chapter, we have briefly introduced the concept of “Data Linkage” and indicate its important role in information extracting semantic structures at different levels. Our contributions to the Data Linkage are outlined. In the following chapter, we will step into understanding the state-of-the-art in Data Linkage and recommend future directions. We will investigate methods which are able to work in different circumstances so that we can improve the effectiveness and efficiency of the data linkage problem.

Chapter 2
A Literature Review of Data Linkage

2.1. Taxonomy of Data Linkage Approaches

Different techniques have been presented by researchers [18, 32, 34, 35, 43, and 77] in multiple areas which argue that the need, task, and type of linkage to be performed will define the involved steps. Other techniques such as the Statistic New Zealand [48] lean toward the idea that data linkage will always require manual preliminary steps such as data classification, sampling and missing observation detection. However, the fundamental problem that arises each time in performing data linkage on large volumes of heterogeneous databases is to discover all possible relationships based on matching similar tuple values that might exist between different table attributes [1].

In this chapter, we survey on techniques that exist in performing approximate data linkage based on their approach rationale. We compare the advantages and disadvantages of current approaches for solving data linkage problem in multiple ways. Our analysis of existing techniques as depicted in Figure 2.1 will show that there is room for substantial improvement within the current state-of-the-art and we recommend techniques where further improvements can be made.

2.1.1. SQL Matching Strategies

SQL Matching techniques [14, 21, 22, 23, 25 and 26] perform data linkage using simple SQL-LIKE commands and SQL Extensions. The advantage of SQL matching techniques is that they help in performing quick data linkage across databases. However, they do not perform well in cases where comparison and identification of data structures need to be performed on large databases containing noisy data without proper unique keys, foreign key relationships, indexes, constraints, triggers, or statistics. Another drawback of the SQL matching process is that it performs |*m*| *x* |*n*| time's column match where *m* and *n* are the total tuple counts in two different databases, resulting in a very slow, expensive and tedious process.

A variation of SQL Matching includes extending query syntax functionalities to perform data linkage. The proposed SQL-LIKE Command languages [22, 23 and 26] handle data transformation, duplicate elimination and cleaning processes supported by regular SQL Query and a proposed execution engine. However, these techniques demand users to have significantly advanced SQL scripting skills and proposed extended functionalities along with sound domain knowledge. Thus, syntax based SQL matching techniques are proven to be less attractive in real world scenarios [22].

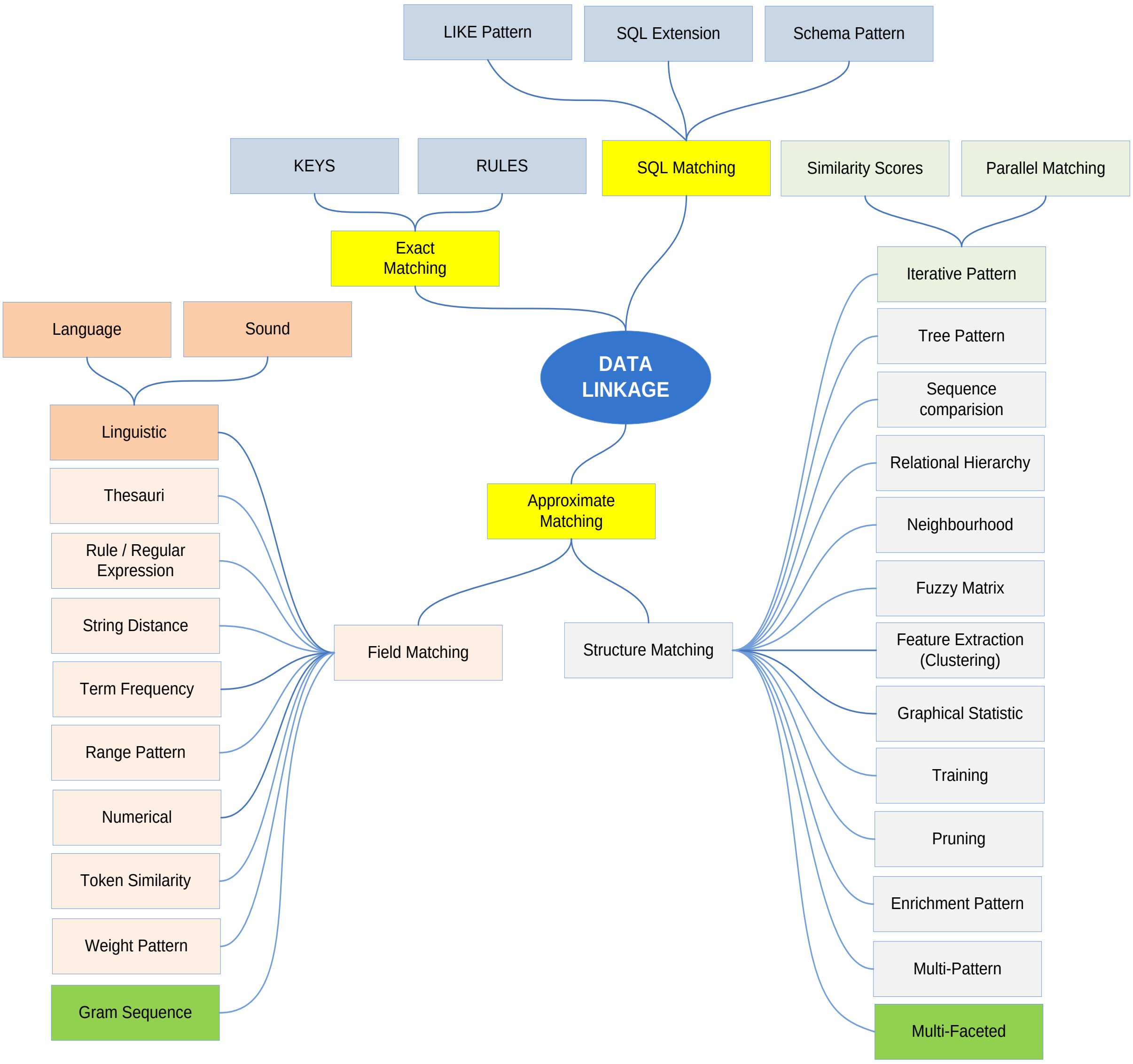


Figure 2.1: Taxonomy of Data Linkage Techniques

Research communities have also stressed Schema Pattern Matching [21 to 26] and SQL Querying [27, 28]. Schema Pattern Matching uses database schema to devise clues as to the semantic meaning of the data. Constraints are used to define requirements, generated by hand or through a variety of tools. However, the main problems with Schema Pattern Matching are insufficiency and redundancy. SQL Querying, on the other hand, uses a SQL query language along with the conceptual modelling extensions such as the Resource Description Framework (RDF) [27, 28] to define matching criteria. Difficulties arise when restrictions eliminate the discovery of possible matches. More relaxed queries use a structure-free mechanism by applying a tree pattern query; however, tree-pattern queries are highly inaccurate due to a high incidence incorrect manual identification of relationships [29].

2.1.2. Exact Matching Strategies

Unlike SQL Matching, Exact Matching techniques give more insight into the content and meaning of schema elements [25]. Exact matching uses a unique identifier present in both datasets being compared. The unique identifier can only be linked to one individual item, or an event (for example, a driver's license number). The Exact Matching technique is helpful in situations where the data linkage to be performed belongs to one data source. For example, consider a company with a recent system crash willing to perform data linkage between the production data source file and the most recent tape backup file to trace transactions. In such situations, Exact Matching would likely suffice in performing data linkage. A specific variation of exact matching discovered In this research is the Squirrel System [31], using a declarative specification language, ISL, to specific matching criteria which will match one record in a given table, with one record in another table. However, exact matching approach leaves no room for uncertainty; records are either classified as a match or as a non-match. Problems often arise when the quality of the variables does not sufficiently guarantee the unique identifier is valid [16]. Exact matching comparison does not suffice for matching records when the data contains errors, for example typographical mistakes, or when the data have multiple representations, such as through the use of abbreviations or synonyms [10].

2.1.3. Approximate Matching Strategies

Approximate Matching is a highly recommended, state-of-the art, alternative approach to exact matching. Approximate matching is also known as the probabilistic approach [34 to 36] within the research community. In approximate matching techniques, data linkage is performed on a likelihood basis (i.e. performing matching based on the success threshold ratio). Output results can vary in different formats such as "match, possible match, and non-match” basis, Boolean type true or false match basis, nearest and outermost distance match basis, discrete or continuous match basis etc. Variations in approximate matching technique include statistical and probabilistic solutions for similarity matching. Attention has also been drawn to approximate matching techniques from different research arenas, including statistical mathematics and bio-medical sciences. Due to the variety of proposed approaches and the level of attributes match, we have further classified approximate matching techniques into attribute level matching and structure level matching groups.

This chapter is organized as follows: In section 2.2, we briefly discuss the multitude of Approximate Matching techniques in the areas of attributes; in section 2.3 we discuss approximate matching approaches at structure level; and in section 2.4 we discuss our conclusions and recommendations for future work.

2.2. Data Linkage: Attribute Level Matching

Attribute Matching, also known as Field Matching [35] and Static String Similarity [36] deals with one-to-one match across different data sources. A challenging task of attribute matching is to perform data linkage across data sources by comparing similar matching records with the assumption that the user is aware of the database structure. Individual record fields are often stored as strings, meaning that functions which accurately measure the similarity of two strings are important for deduplication [36]. In the following subsections, we describe attribute matching methodologies and discuss the efficiency of each.

2.2.1. Linguistic similarity

Linguistic techniques focus on phonetic similarities between strings. The rationale behind this approach is that while strings may be similar phonetically, they may have different characters to locate potential matches. Soundex [34] is the most widely known in this area, and uses codes to define letters, remaining non-coded letters are used as separators. In addition, Soundex checks for identical codes (A, E, I, O, U and Y) without separators. Through the Soundex rules, a possible match is determined or denied. Advantages of linguistic techniques include the exposure of about 2/3 of spelling variations [25, 32, and 34]. However, linguistic methods are not equally effective from one ethnicity to the next. Linguistic based techniques are designed for Caucasians, and works on most other ethnicities, but largely fails on East Asian names due to the phonetic differences. NYSIIS improves upon this by maintaining vowel placement and converting all vowels to the letter A. Nonetheless, it is still not perfectly accurate and performs best on surnames and not on other types of data [34].

2.2.2. Rule/Regular expression

The Rule / Regular expression [40] approach uses rules or set of predefined regular expressions and perform matching on tuples. Regular Expression Pattern as proposed in [40] is more flexible than regular expression alone, which is built from alphabetical elements. This is also because the Regular Expression Pattern is built from patterns over a data element, allowing the use of constructs such as “wildcards” or pattern variables. Regular Expression Pattern is quite useful when manipulating strings, and can be used in conjunction with basic pattern matching. However, the problem with this approach lies in the fact that it is relatively domain specific and tends to only work well on strings.

2.2.3. Ranking

Ranking [15, 41] methods determine preferential relationships and have been more recently recognized by researchers as a necessary addition to structure based matching techniques. Search engines have used ranking methods for some time, such as Google’s PageRank, despite such algorithms not suited for matching noisy data due to their poor connectivity and lack of referrals [15]. Therefore, ranking extensions which simultaneously calculate meaning and relevance are researched. Thus far, only a few ranking methods have been proposed including induction logic programming, probabilistic relational kernel, and complex objects ranking [15, 41].

2.2.4. String distance

String distance methods, also known as character-based similarity metrics [34] are used to perform data linkage based on the cost associated within the comparing strings. The cost is estimated on the number of characters which needs to be inserted, replaced or deleted for a possible string match. For example, Figure 2.2 shows the cost associated in editing string “Aussie” to “Australian” (the “+” sign shows addition, the “-“ sign shows deletion, and the “x” sign shows replacement):

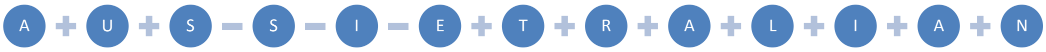


Figure 2.2: This is a simple example of string distance technique for editing string “Aussie” to “Australian”.

Experimental results in [34] have shown that the different distance based methodologies discovered so far are efficient under different circumstances. Some of the commonly recommended distance based metrics include Levenstein distance, Needleman-Wunsch distance, Smith-Waterman distance, Affine-gap distance, Jaro metric, Jaro and Jaro-Winkler metric, Q-gram distance, and positional Q-grams distance. Through the various methods, costs are assigned to compensate for pitfalls in the system. Yet, overall, string distance pattern is most effective for typographical errors, but is hardly useful outside of this area [34].

2.2.5. Term frequency

Term frequency [43] approach determines the frequency of strings in relation and to favour matches of less common strings, and penalizes more common strings. The Term frequency methods allow for more commonly used strings to be left out of the similarity equation. TF-IDF [43] (Term Frequency-Inverse Document Frequency) is a method using the commonality of the term (TF) along with the overall importance of the term (IDF). TF-IDF is commonly used in conjunction with cosine similarity in the vector space model. Soft TF-IDG [44] adds similar token pairs to the cosine similarity computation. According to the researchers in [44], TF-IDF can be useful for similarity computations due to its ability to give proportionate token weights. However, this approach fails to make distinctions between the similarity level of two records with the same token or weight, and is essentially unable to determine which record is more relevant.

2.2.6. Range pattern

Range pattern matching returns a Boolean style true or false result if the specified tuples fall within the specified range. Similarity or dissimilarity is determined when the elements of the data are compared against the predetermined range. Range matching will return a 0 or 1, with 0 being false and 1 being true. Range pattern matching is often used as an expansion of an algorithm to filter results. For example, TeenyLIME [45] expands upon LIME by adding range pattern capabilities, giving TeenyLIME the ability to define the range of its results. A drawback of the range pattern approach is that it is often not powerful enough to perform matching without a high level of query knowledge. For example, if a query is made to search for nearby locations, an optimal range is often not given or is defined by words having various meanings, causing range pattern matching to produce inaccurate results.

2.2.7. Numeric distance

Numeric distance methods are used to quickly perform data linkage on tuples that contains numerical values but don't require complex string character-style comparison. Hamming distance [46], for example, is used for numeric values such as zip codes, and counts the variations between two records. Due to the limitations of numeric data type constraints, it has not received much attention. Numeric distance methods can be best used in combination of other techniques.

2.2.8. Token matching

Token based matching compare fields by ignoring the ordering of the tokens (words) within these fields. Token based approach use tokenization to perform matching, which is the separation of strings into a series of tokens. It assigns a token to each word in the string and tries to perform matching by ignoring token order and by performing similar match. The token based approach attempts to compensate for the inadequacies of character-based metrics, specifically the inability to detect word order arrangement. A tokenizer performs the operation, taking into account characters, punctuation marks, blank spaces, numbers, and capitalisation. Token based methods count a string as a word set, and accommodates duplicates. For example, Cosine Similarity [38] is used to perform data linkage based on record strings, irrespective of word ordering within the string. The Cosine Similarity methods are effective over a range of entry types, and also have the advantage of considering word location to allow for swapping of word positions. For data containing a large amount of text, the token based matching works quite well, as it can handle repeating words. The optimising token based approach has typically included aggregation of different sources. A potential drawback is that token based matching does not store substring order and can predict false matches.

2.2.9. Weight pattern

Weight pattern also referred to as Scoring [47], is applied on matching strings to return a numerical weight; a positive weight for agreeing values and a negative weight for disagreeing values. As two records are compared, the system assigns a weight value for similarity comparison. Composite weight [48] is a summation of all the field weights for a record, which multiplies the probabilities of each value. Reliability of the information, commonality of the values, and similarity between the values are considered in determining weight. Determinations are made by calculating the “m” probability (reliability of data) and the “u” probability (the commonness of the data). For example, IDF weights consider how often a particular value is used. After weights are determined for all the data, cut-off thresholds are set to determine the comparison range. Unfortunately, weight pattern techniques do not perform well when there are data inconsistencies. True matches may have low weights, and non-matches may have high weights as a result of simple data errors [48].

2.2.10. Gram sequence

Gram sequence based techniques compare the sequence of grams of one string with the sequence of grams of another string. n-grams is a gram based comparison function which calculates the common characters in a sequence, but is only effective for strings that have a small number of missing characters [46]. For example, the strings “Uni” and “University” have the same 2-gram {un, ni}. q-gram [85] involves generating short substrings of length *q* using a sliding window at the beginning and end of a string [85]. The q-gram method can be used in corporate databases without making any significant changes to the database itself [85]. Theoretically, two similar strings will share multiple q-grams. Positional q-grams record the position of q-grams within the string [14]. Danish and Ahy in [85] proposed to generate q-grams along with various processing methods such as substrings, joins, and distance. Unfortunately, the gram sequence approach is only efficient for short string comparison and becomes complex, expensive and unfeasible for large strings [85].

2.2.11. Blocking

Blocking [46] techniques separate tuple values into set of blocks/groups. Within each of these blocks, comparisons are made. Sorted Neighborhood is a blocking method which first sorts and then slides a “window” over the data to make comparisons [46]. BigMatch [51] used by the U.S. Census Bureau, is another blocking technique. BigMatch identifies pairs for further processing through a more sophisticated means. The blocking function assigns a category for each record and identical records are given the same category. The disadvantage of the blocking method is that it will not work for records which have not been given the same category [18, 25, and 34].

2.2.12. Hashing

Hashing methods convert attributes into a sequence of hash values which are compared for similarity matching between different sets of strings. Hashing methods require conversion of all the data to find the smallest hash value, which could be a costly approach. Set-of-sets [8] is a hashing based data matching technique which works reasonably well in smaller string matching scenarios. The set-of-sets technique proposed in [8] divides strings into 3-grams and assigns a hash value to each tri-gram. Once hash values are assigned and placed in a hash bag, only the lowest matching hash values are considered for matching. Unfortunately, this technique doesn’t yield accurate results when dealing with variable length strings and uses traditional hashing which results in completely different hash values for even a small variation [79]. Furthermore, the Set-of-sets requires conversion of all the data prior to comparison in order to find the smallest hash value, which could be a costly approach. To overcome this disadvantage, the h-gram (hash gram) method was proposed in [79] to address the deficits of the set-of-sets technique, by extending the n-gram technique; utilizing scale based hashing; increasing matching probability; and by reducing the cost associated in storage of hash codes.

2.2.13. Path sequence

The path sequence approach such as in [37] examines the label sequences, and compares them to the labelled data. The distance is measured by determining the similarity between the last elements of a path. The prefix can be considered, but this only affects the result to a certain degree, and becomes less relevant with increasing distance between the prefix and the end of the sequence.

2.2.14. Conditional substrings

Substring matching such as in [53] expands upon string-based techniques by adding substring conditions to string algorithms. Distance measurements are calculated for the specified substring, in which all substring elements must satisfy the distance threshold. A frequent complication related to conditional substring based matching involves the estimation of the size of intersection among related substrings. Clusters and q-grams, which are commonly used in string estimation, are not applicable in substring based techniques, because substring elements are often dissimilar. As a result, substring matching is hindered by an abundance of possibilities, which must all be considered.

2.2.15. Fuzzy matrix

Fuzzy Matrix [32, 60] places records in the form of matrices and apply fuzzy matching techniques to perform record matching. Commonly used by social scientists to analyse behavioural data, the fuzzy matrix technique is also applicable to many other data types. When considering a fuzzy set, a match is not directly identified as positive or negative. Instead, the match is considered on its degree level of agreement with the relevant data. As a result, a spectrum is created which identifies all levels of agreement or truth.

2.2.16. Thesauri matching

Thesauri based matching attempts to integrate two or more thesauruses. A thesaurus is a kind of lexicon to which some relational information has been added, containing hyponyms which give more specific conceptual meaning. WordNet [27, 32, and 52] is a public domain lexical database, or thesaurus, which makes its distinctions by grouping words into sets of synonyms; it is often used in thesauri matching techniques. Falcon and DSSim [52] are thesauri based matching tools which incorporate lexicons, edit-distance and data structures. LOM [32] is a lexicon-based mapping technique using four methods (whole term, word constituent, synset, and type matching) in an attempt to reduce the required amount of human labor, but does not guarantee any level of accuracy. While Thesauri based approaches can be extremely useful in merging conceptual, highly descriptive information; they can be incredibly complex and difficult to automate to a significant degree; and human experts are typically required to quality assure the relationships [27]. Thesauri matching algorithms must consider the best balance between precision and recall.

2.3. Data Linkage: Structure Level Matching

Structure level matching is used when the records being matched need to be fetched from a combination of records (i.e. when attempting to match noisy tuples across different domains, and requiring more than one match). Grouped attribute matching techniques perform data matching, with the main intuition that the grouping of attributes into clusters followed by performing matching provides a deeper analysis of related content and semantic structure. This process was initially considered for discovering candidate keys and dependent keys. However, one of the biggest challenges involved in this process has been the large number of combinations required for grouping attributes and performing data matching between these groups, which can be costly and time consuming [25, 32, 34 and

37]. Large scale organisations such as Microsoft and IBM have introduced Performance Tuner tools for indexing combined attributes on which queries are frequently executed. Unfortunately, these tools are suited to Database Developers / DBA's who have sound knowledge in executing SQL queries and is not ideal for novice users. As such, research has taken new direction by classifying multiple structure level techniques that require matching across multiple attributes. We have classified these techniques in the following subsections.

2.3.1. Iterative pattern

Iterative pattern is the process of repeating a step multiple times (or making "passes") until a match is found based on similarity scores and blocking variables (variables set to be ignored for similarity comparison). The Iterative approach uses attribute similarity, while considering the similarity between currently linked objects. For example, the Iterative pattern method will consider a match of "John Doe" and "Jonathan Doe" as a higher probability if there is additional matching information between the two records (such as a spouse's name and children's names). The first part of the process is to measure string distance, followed by a clustering process. Iterative pattern methods have proven to detect duplicates that would have likely been missed by other methods [54]. The gains are greater when the mean size of the group is larger, and smaller when the mean size is smaller. Disadvantages surface when distinctive cliques do not exist for the entities or if references for each group appear randomly. Additionally, there is also the disadvantage of cost, as the Iterative pattern method is computationally quite expensive [54].

2.3.2. Tree pattern

Tree pattern is based on decision trees with ordered branches and leaves. The nodes are compared based on the extracted tree information. CART and C.5 are two widely-known decision tree methods which create trees through an extensive search of the available variables and splitting values [55]. A Tree pattern starts at the root node and recursively partitions the records into each node of the tree and creates a child to represent each partition. The process of splitting into partitions is determined by the values of some attributes, known as splitting attributes, which are chosen based on various criteria. The algorithm stops when there are no further splits to be made. Hierarchical verification through trees examines the parent once a matching leaf is identified. If no match is found within the parent, the process stops; otherwise the algorithm continues to examine the grandparent and further up the tree [37]. Suffix trees such as DAWG [37] build the tree structure over the suffixes of S, with each leaf representing one suffix and each internal node representing one unique substring of S. DAWG has additional feature of failure links added in for those letters which are not in the tree. Disadvantages of Tree pattern lies in lengthy and time consuming process with manual criteria often needed for splitting.

2.3.3. Sequence pattern

Sequence pattern methods perform data linkage based on sequence alignment. This technique attempts to simulate a sequential alignment algorithm, such as the BLAST (Basic Local Alignment Search Tool) [12] technique used in Biology. The researchers compared the data linkage problem with the gene sequence alignment problem for pattern matching, with the main motivation to use already invented BLAST tools and techniques. The algorithm translates record string data into DNA sequences, while considering the relative importance of tokens in the string data [12]. Further research in the Sequence pattern area have exposed variations based on the type of translation used to translate strings into DNA Sequence (i.e. weighted, hybrid, and multi-bit BLASTed linkage) [12]. BLASTed linkage has advantages through the careful selection of one of its four variations, as each variation performs well on specific types of data. Unfortunately, sequence pattern tends to perform poorly on particular data strings, depending upon the error rate, importance weight, and number of common tokens [12].

2.3.4. Neighbourhood pattern

The neighbourhood approach [7, 59] attempts to understand and measure distribution according to their pattern match, and is a primary component in identifying statistical patterns. By using the nearest neighbour approach, related data is able to be clustered even if it is specifically separated. The logic behind this approach is based on the assumption that, if clustered objects are similar, then the neighbours of clustered objects have a higher likelihood of also being similar. Neighbourhood pattern requires a number of factors that need to be carefully considered in order to determine pattern matches.

2.3.5. Relational hierarchy

Relational Hierarchy techniques use primary and foreign key relationships to understand related table content in order to perform data linkage. Relational hierarchy forms relation links which connect concepts within various categories. It breaks down the hierarchical structure and the top-level structure contains children sets. The relational hierarchy technique compares and calculates the co-occurrence between tuples by measuring the overlap of the children sets. A high degree of overlap will indicate a possible relationship between the two top level categories [57]. Relational Hierarchy techniques are only effective when primary and foreign key relationships have been established. Raw data, without predefined relationships, cannot be linked using this approach.

2.3.6. Clustering / Feature extraction

Clustering, also known as the Feature extraction method performs data linkage based on common matching criteria in clusters, so that objects in clusters are similar. Soft clustering [61], or probabilistic clustering, is a relaxed version of clustering which uses partial assignment of a cluster center. The SWOOSH [62] algorithms apply ICAR properties (idempotence, commutativity, associativity, representativity) to the match and merge function. With these properties and several assumptions, researchers introduced the brute force algorithm (BFA), including the G, R and F SWOOSH algorithms [44]. SIMCLUST is another similarity based clustering algorithm which places each table in its own cluster as a starting point and then works its way through all of the tables by consecutively choosing two tables (clusters) with the highest level of similarities. [5] proposed iDisc system which creates database representations through a multi-process learning technique. Base clusters are used to uncover topical clusters which are then aggregated through meta-clustering. Clustering in general can get extremely complex (such as forming clusters using semantics) and needs to be handled carefully while discovering relationships between matching clusters.

2.3.7. Graphical statistic

Graphical statistic is a semi-automated analysis based technique where data linkage is performed based on the results obtained on the graph. Such representations illustrate the topical database structure through tables. The referential relationship indicates an important linkage between two separate tables. Foreign keys within one table may refer to keys within the second table. However, problems with this technique often arise due to the fact that information on foreign keys is often missing [5].

2.3.8. Training based

Training based technique is a manual approach where users are constantly involved in providing statistical data based on previous/future predictions. In [7], researchers presented a two-step training approach using automatically selected, high quality examples which are then used to train a support vector machine classifier. The approach proposed in [7] outperforms k-means clustering, as well as other unsupervised methods. the Hidden Markov training model, or HMM, standardises name and address data as an alternative method to rule-based matching. Through use of lexicon-based tokenization and probabilistic hidden Markov models, the approach attempts to cut down on the heavy computing investment required by rule programming [64]. Once trained, the HMM can determine which sequence of hidden states was most likely to have emitted the observed sequence of symbols. When this is identified, the hidden states can be associated with words from the original input string. This approach seems advantageous in that it cuts down on time costs when compared to rule-based systems. However, this approach remains a lengthy process, and has shown to run into significant problems in various areas. For instance, HMM confuses given, middle, and surnames, especially when applied to homogenous data. Furthermore, outcomes proved to be less accurate than those of rule-based systems [64]. DATAMOLD [65] is a training-based method which enhances HMM. The program is seeded with a set of training examples which allows the system to extract data matches. A common problem with training techniques is that it requires many examples to be effective; and the system will not perform without an adequate training set [55].

2.3.9. Pruning / Filtering statistic

Pruning statistic performs data linkage by trimming similar records on a top down approach. In [16], the data cleaning process of “deduplication” involves detecting and eliminating duplicate records to reduce confusion in the matching process. For data which accepts a large number of duplicates, pruning, before data matching, simplifies the process and makes it more effective. A pruning technique proposed by Verykios [34] recommends pruning as on derived decision trees used for classification of matched or mismatched pairs. The pruning function reduces the size of the trees, improving accuracy and speed [34]. The pruning phase of CORDS [16] (which is further discussed in the statistical analysis section) prunes non-candidates on the basis of data type, properties, pairing rules, and workload; such tasks are done to reduce the search space and make the process faster for large datasets. Pruning techniques [37] are based on the idea that it is much faster to determine non-matching records than matching records, and therefore aim to eliminate all non-matching records which do not contain errors. However, the disadvantage of such techniques is that they are not suitable in identifying matches of any type, and must be combined with another matching technique.

2.3.10. Enrichment pattern

Enrichment patterns are a continuous improvement based technique which performs data linkage by enriching the similarity tasks on a case by case basis. An example of the enrichment method is ALIAS [34], a learning-based system, designed to reduce the required amount of training material through the use of a “reject region”. Only pairs with a high level of uncertainty require labels. A method similar to ALIAS is created using decision trees to teach rule matching in [34]. OMEN [32] enriches data quality through the use of a Bayesian Net, which uses a rule set to show related mappings. Semantic Enrichment [66] is the annotation of text within a document by sematic metadata, essentially allowing free text to be converted into a knowledge database through data extraction and data linking. Conversion to a knowledge database can be through exact matching or by building hierarchical classifications of terms; text mining techniques allow annotation of concepts within documents which are subsequently linked to additional databases. Thesauri alignment [32, 52] based techniques are also considered as part of enrichment techniques because it combines concepts and better defines the data. The problems associated with enrichment approach include substantial investment of time and the requirement for extensive domain knowledge.

2.3.11. Multiple pattern

The multiple pattern approach performs data linkage through the simultaneous usage of different matching techniques. This approach best fits when one does not know which technique performs better. The researchers in [31] use a multi approach which combines sequence matching, merging, and then exact matching. Febrl [67] is an open-source software containing comparison, and record pair classifications. Febrl results are conveniently presented in a graphical user interface which allows the user to experiment with numerous other methods [67]. TAILOR [46] is another example which uses three different methods to classify records: decision tree induction, unsupervised k-means clustering, and a hybrid approach. GLUE [68] is yet another matching technique allowing for multiple matching methods. GLUE performs matching by first identifying the most similar concepts. Once these concepts are identified, a multi-strategy learning approach allows user to choose from several similarity measures to perform the measurement. In our research, we have provided an extended multi-strategy approach through introducing Multi-Modular Neural Network [1, 79, and 106]; an ontology based learning approach for categorizing given data into predefined classes, based on similarities in their ontologies.

2.3.12. Data constraints

Data constraints, also known as internal structure based techniques, apply a data constraint filter to identify possible matches [43]. The constraint typically uses specific criteria of the data properties. This technique is not suited when used on its own, and performs best for the elimination of non-matches, as a pre-processing method before a secondary method, such as clustering. Furthermore, data constraints don't handle the large number of uncertainties present within the data. Hence, adding constraints for each uncertainty is computationally infeasible.

2.3.13. Taxonomy

Taxonomy based methods use taxonomies, a core aspect of structural concepts which are largely used in file systems and in knowledge repositories [69]. This approach uses the nodes of taxonomy to define a parent/child relationship within the conceptual information and create classification. Using specified data constraints, the taxonomy of multiple data sources are evaluated into a technique known as structural similarity measure. For example, in [70] researchers used a taxonomy mapping strategy to enrich WordNet with a large number of instances from Wikipedia, essentially merging the conceptual information from the two sources. As with similar methods, taxonomy based matching requires a significant degree of domain knowledge and performs with limited precision and inadequate recall.

2.3.14. Hybrid match

Hybrid techniques use a combination of several mapping methods to perform data match. A prime example of the hybrid method is described in [71], which uses a combination of syntactic and semantic comparisons. The rationale behind hybrid matching is that the semantics alone is not sufficient to perform accurate matching and could be inconsistent. The hybrid solution consists of a hybrid of semantic and syntactic matching algorithms which considers individual components. The syntactic match uses a similarity score based on class, prefix and substring, and the semantic match uses a similarity score based on cognitive measures such as LSA, Gloss Vector, and WordNet Vector. The information is aggregated and entered into a matrix and experts are used to determine domains within the selected threshold.

2.3.15. Data extraction

Data extraction primarily involves extracting semantic data. Data extraction can be performed manually or with an induction and automatic extraction [72]. In [73], researchers used data recognisers to perform data extraction on the semantics of data. The recogniser method is aimed at reducing alignment after extraction, speeding up the extraction process, reusing existing knowledge, and cutting down on manual structure creation. This approach is found to be effective for simple unified domains, but not for complicated, loosely unified domains. Another benefit of the data extraction technique is that, after the data is extracted, it can be handled as instances in a traditional database. However, it generally requires a carefully constructed extraction plan by an expert in that specific knowledge domain [74].

2.3.16. Knowledge integration

Knowledge integration techniques are used to enhance the functioning of structure level matching by integrating knowledge between data relationships to form a stronger concept base for performing data linkage [75]. Knowledge integration enhances query formulation when the information structure and data sources are not known, as highlighted in [76], and is becoming increasingly important in data matching processes as various data structures conceptualise the same concept in different ways, with resulting inconsistencies and overlapping material. Integration can be based on extensions or concepts, and is aimed at indemnifying inconsistencies and mismatches in the concepts. For example, the COIN technique [77] addresses data-level heterogeneities among data sources expressed in terms of context axioms and provides a comprehensive approach to knowledge integration. An extension of COIN is ECOIN, which improves upon COIN through its ability to handle both data-level and ontological heterogeneities in a single framework [77]. Knowledge integration is highly useful in medicine, to integrate concepts and information within various medical data sources. Knowledge integration involves the introduction of a dictionary to fill knowledge gaps, such as using distance-based weight measurement through Google [68]. For example, the Foundational Model of Anatomy is used as a concept roadmap to better integrate various medical data sources into unique anatomy concepts [68].

2.3.17. Data structures

Data structures use structural information to identify match and reflect relationships. Information properties are often considered and compared with concepts to make a similarity determination, while other variations of the data structure approach uses graphical information to create similarities [68]. A drawback of the data structure based approach results from its consumption rate of resources; the process builds an “in-memory” graph containing paired concepts which can lead to memory overflow.

2.3.18. Statistical analysis

Statistical analysis techniques examine statistical measurements for determining term and concept relationships. Jaccard Similarity Coefficient [38] is a widely used statistical measurement for comparing terms, which consider the extent of overlap between two vectors. The measurement is the size of the intersection, divided by the size of the union of the vector dimension sets. Considering the corpus, the Jaccard Similarity approach determines a match to be present if there is a high probability for both concepts to be present within the same section. For attribute matching, a match is determined if there is a large amount of overlap between values [38]. For example, CORDS [16] is a statistical matching tool, built upon B-HUNT, which locates statistical correlations and soft functional dependencies. CORDS searches for correlated column pairs through enumerating potentially correlating pairs and pruning unqualified pairs. A chi-squared analysis is performed in order to locate numerical and categorical correlations. Unfortunately, statistical analysis methods are generally restricted to column pairs, and may not detect correlations where not all subsets have been correlated [1, 18].

2.4. Summary

The data linkage approaches reviewed in this chapter represents a variety of linkage techniques using different aspects of data. We discussed practical methods from two different angles, they are, Attribute level and Structure-level approaches. We showed that classification of data into a single order does not provide the necessary flexibility for accurately defining data relationships. Furthermore, we found that the flow of data and their relationships need not be in a fixed direction. This is because, when dealing with variable data sources, same sets of data can be ordered in multiple ways based on the semantics of tables, attributes and tuples. This is critical when performing data linkage. We proposed that one of the most promising approaches which can further be developed is the scale based hashing [79, 106] technique, as we see the uniformity of hashing as a base point for the development of a globally applicable hash code system.

Through our analysis of the status quo we proved that the research should take a new direction to discover possible data matches, based on its inherent hierarchical semantic similarities. This approach is ideal for knowledge based data matching and query answering. We recommend faceted classification to classify data in multiple ways, to source semantic information for accurate data linkage and other data intrinsic tasks. We recommend, in response to the intricacy of this background research, that the data linkage research community collaborate to benchmark existing data linkage techniques, as it is getting increasingly complicated to convincingly and in a timely manner compare new techniques with existing ones.

In chapter 3, we will first deal with the problem of reducing the number of comparisons required for the data linkage process on a variety of data types and data sizes at various attribute levels. We will formally introduce our new hash gram (h-gram) record matching technique to deal with this problem.

In chapter 4, we will consider the research problem of constructing a ‘knowledge based’ multi-faceted cluster mapping technique, which aims at extracting probable relationships between correlated data clusters on a structure level. We will formally introduce the linkage problem through our Ontology Guided Data Linkage (OGDL) framework and show how our algorithms can be applied to heterogeneous databases.

Chapter 3
Approximate Record Matching Using Hash Grams

3.1. Data Translation

The central problem that arises each time when attempting to link heterogeneous databases is to perform Record Matching. Obviously, Record Matching is not a new issue. Efficient and accurate Record Matching (also referred to as Data Matching, Instance Identification, Record Linkage, De-duplication, Data Cleaning, Entity-Resolution and Merge Purge) has been a well-known problem within the research community [2, 34]. Due to its significant demand as such, much progress has been made in finding different logical and statistical ways to solve linkage problems. At the core of this issue lies in performing one-to-one variable record matching. The computational expense derived from performing such a pair-wise record matching has been the main drawback of existing techniques especially when dealing with noisy data [5, 8, 14 and 18]. Approximate Matching is a highly recommended, state-of-the art, alternative approach to exact matching [34 to 36]. Systematic engineering has emerged in recent years to build tools that use modern technologies with full, semi-automated or controlled based approaches, depending on the need and the area of research they are working on. Unfortunately, when attempting to perform record matching where there are inconsistencies in the data, implementation of these techniques is highly expensive, time consuming and limited to specific data spaces, without support for ad-hoc record matching [1].

In conducting our research, we investigated the record-matching problem by analysing and performing experiments on thousands of real world data from a variety of sources [24, 30, 33, 42, 49, 50, and 56], anticipating that the lack of a common domain and having inconsistencies in data would effectively address the shortage of multi-domain experimentation in current research. An effective way to matching similar records is to transform given raw-data obtained from heterogeneous data sources into its equivalent measurable units or numerical facts. The advantage of such a transformation is to deal with record linkage problem by using a statistical approach. Any transformation technique we choose should ensure that it can accept any given data type. Moreover, in order to perform probabilistic data matching, the transformation should guarantee the accuracy, performance and correctness and should preserve the data essence.

In this chapter, we are going to introduce scale based gram hashing technique in order to assign and associate meaningful numerical equivalent hash codes to each data element (gram) that can assist in identifying similar data across multiple data sources. The rationale behind this approach is that “common hash codes have to be associated with common strings”, or in other words, “the similar records should have similar hash codes”. This technique overcomes the disadvantages of random-assignation hashing systems. Furthermore, the costs associated in running existing techniques are comparatively high when aiming at performing ad-hoc integration and to get approximate results in a timely manner. The proposed h-gram technique overcomes the disadvantages of existing techniques and performs record matching in a quick and dirty process highly reducing the runtime cost and providing a way of getting approximate and reliable record matching results.

3.1.1. Problem Statement

Hashing is the process of assigning of numerical values (hash codes) to data and subsequently categorizing the data by the assigned code. Traditional hashing creates unique values for each variation in a string. In other words, traditional hashing will result in a completely different number sequence for even a small variation. For example, traditional hashing will assign a hash code of 231082007 for “invoice” and 1218906135 for “invoices” despite strings being nearly identical in spelling and meaning. As a result, traditional hash functions can produce an unmanageable number of values, rendering the method far too complex for universal use.

In order to reach the creating describable aggregations from sets of heterogeneous raw data inputs, non-scaled entries (strings or sets of characters) must be converted into numerically-significant realities i.e. hashing the input entries before performing data matching [8, 18]. The resulting hash codes (numerical facts) must preserve the original string’s essence, meaning that the results of comparing the hash codes should resemble the results of comparing the strings. In order to reduce the computationally expensive process of generating hash codes by preserving the string case, our proposed technique will treat any given data as non-case sensitive. This assumption highly reduces the associated time and cost in order to get probabilistic results.

3.1.2. Our Approach

N-grams/Q-gram [34, 85] techniques calculate the distance between two substrings defined by a length of n. The current problem with traditional n-gram lies in the abundance of hash values generated for each gram, making it extremely difficult to perform matches (or likely matches) in a timely manner [1, 2]. Higher values of n will yield more possible matches while lower values of n will yield increasingly fewer possible matches. Bigrams, where n = 2 has been used to calculate small spelling errors between two otherwise identical strings. Trigrams, where n = 3 has been used to identify duplicate records in bibliographic records. The methods of n-grams are used in our algorithm to scale the level of detail in data analysis.

Standard blocking techniques [5, 88, and 90] separate data into categories called blocks or ‘buckets’. Data comparisons are only made between records that fall within the same block. Another example of blocking is the Sorted Neighbourhood [7, 59] approach which initially sorts data, then follows with a sliding window technique that slides a predefined “window” over a set amount of characters for the records that fall within the given block. The downside of these techniques is that they are costly when dealing with noisy data and are not efficient for retrieving probabilistic results [2, 79]. To overcome this problem, in our research, we employed probabilistic sliding window while generating grams and splitting larger strings into sub-string comparisons.

Set-of-sets [8] is a data matching technique which works reasonably well in smaller string matching scenarios. The set-of-sets technique divides strings into 3-grams and assigns a hash value to each tri-gram. Once hash values are assigned and placed in the hash bag, only the lowest matching hash values are considered for matching. Unfortunately, our experimental results (see section 3.6) have shown that this technique doesn’t yield accurate results when dealing with variable length strings and uses traditional hashing which results in creating completely different hash values for even a small variation. Furthermore, the Set-of-sets requires conversion of all the data to find the smallest hash value which could be a costly approach.

Our proposed h-gram method will address the deficits in the set-of-sets [8] technique by extending n-gram method; utilizing scale based hashing; increasing matching probability; and finally reducing the cost associated in storage of hash codes.

3.2. Data Transformation

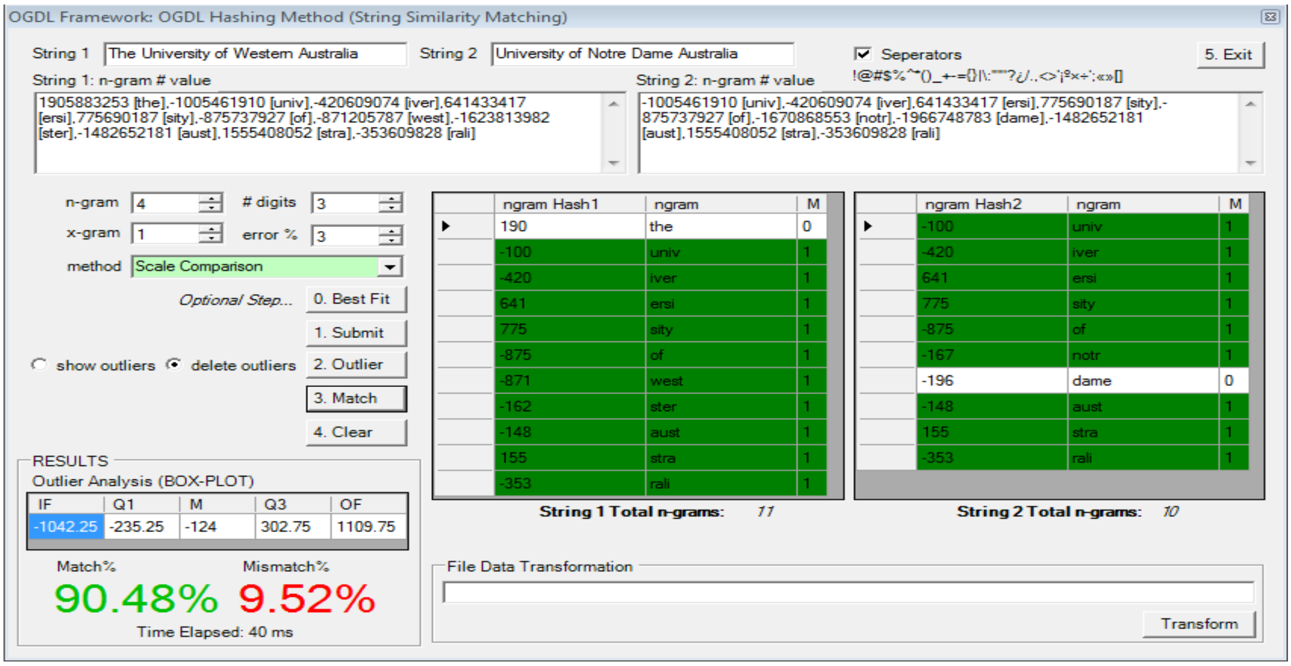
Data Transformation is the process of transforming the given raw-data obtained from heterogeneous databases into its measurable units or numerical facts. This technique becomes extremely important especially when analyzing large volumes of string data type attributes to reduce associated time and costs. Hence, performing data linkage in the collected samples from heterogeneous databases can be a complex process if we don’t have statistical measurements in place. Any transformation technique we choose should ensure that it can accept any given data types.

Generating hash codes is a well-known process aimed at quickly transforming hash values for equality testing. The best way to meet the intended target (similar hash codes for similar words) independently upon the inputted string, its number of characters and the location of their constituent tuples (“tri” in “tricycle” and in “geriatric”), is generating hash codes where both issues (number of characters and their position) are clearly identified. In other words, the number of characters and the location of their constituent tuples must be considered. In order to represent each gram value into its constituent numerical fact for similarity comparison, we transform string gram using scale based hashing such as [20] and considering only the first *n* number of hash digit values for probabilistic matching. Alternatively, any kind of hashing method can be used as long as it meets the main point of purpose i.e. to return similar hash codes for similar strings. Table 3.1 lists examples of data transformation process performed on different strings with μ = 3 (where μ is the first *n* digits for consideration, set by the user)

Table 3.1: Sample hash codes comparison using h-gram

-1505962632 {john}	≈	-1505634957 {ojhn}
-1100074170 {meke}	≈	-1099549882 {mike}
-332410181 {ngram}	≈	-332409375 {xgram}
405195713 {4/11}	≈	405195712 {5/11}
133772777 {caf}	≈	133903849 {cof}
1699529035 {ffe}	≈	1699529035 {ffe}

The data transformation process is aimed at performing approximate record matching and we should expect a level of false positive matches with this technique. Nevertheless, the numerical realities obtained through our data transformation process when applied on large samples of data are quiet high and from the data linkage point of view, this technique highly reduces the number of comparisons required as detailed in the next section.



String 1: The University of Western Australia

String 2: University of Notre Dame Australia

☑ Seperators !@#$%^~()_+-={}|\:""""?¿/.,<>'¡º×÷';«»[]

String 1: n-gram # value

1905883253 [the],-1005461910 [univ],-420609074 [iver],641433417 [ersi],775690187 [sity],-875737927 [of],-871205787 [west],-1623813982 [ster],-1482652181 [aust],1555408052 [stra],-353609828 [rali]

String 2: n-gram # value

-1005461910 [univ],-420609074 [iver],641433417 [ersi],775690187 [sity],-875737927 [of],-1670868553 [notr],-1966748783 [dame],-1482652181 [aust],1555408052 [stra],-353609828 [rali]

n-gram: 4 # digits: 3

x-gram: 1 error %: 3

method: Scale Comparison

○ show outliers ● delete outliers

ngram Hash1	ngram	M
190	the	0
-100	univ	1
-420	iver	1
641	ersi	1
775	sity	1
-875	of	1
-871	west	1
-162	ster	1
-148	aust	1
155	stra	1
-353	rali	1

String 1 Total n-grams: 11

ngram Hash2	ngram	M
-100	univ	1
-420	iver	1
641	ersi	1
775	sity	1
-875	of	1
-167	notr	1
-196	dame	0
-148	aust	1
155	stra	1
-353	rali	1

String 2 Total n-grams: 10

RESULTS

Outlier Analysis (BOX-PLOT)

IF	Q1	M	Q3	OF
-1042.25	-235.25	-124	302.75	1109.75

Match% 90.48% Mismatch% 9.52%

Time Elapsed: 40 ms

File Data Transformation

Figure 3.1: h-gram record matching prototype

3.3. Definitions and Notations

The h-gram matching algorithm (see Algorithm 1) considers the estimation of a family of parameters for any given set of transformed data. In this section, we provide a detailed description of each of these parameters which can be configured by the ultimate user at run time in order to generate sets of hash grams as depicted in Figure 3.1. Given a set of transformed values, the h-gram algorithm applies these parameter settings and performs flexible iterations until the desired precision of the estimated matches are obtained.

3.3.1. String sets

String-sets ΣSs_n are the substrings the application will generate in sequence in order to split larger strings into an array of sub-strings using the given set of separators. The strings are considered for splitting into substrings only if the string exceeds the maximum length γ (value set by the user). This step is required in order to reduce the cost associated in performing one-to-one long string matches at the same time preserving the string's sequential essence. Suppose we have $w_{1,}w_2, w_{3,}\ and \dots\ w_n$ ordered words in a string S_1. Let $(s_1,\ s_2, \dots s_n)$ be the given set of separators. Let the minimum words count γ = 25. Then, the String-sets ΣSs_n are created by splitting into sets of sub-strings.

$$\Sigma Ss := \{(w_{1,}w_2, \dots\ w_n) \dots . (w_{1,}w_2, \dots\ w_n) \mid (s_1,\ s_2, \dots s_n)\}$$

(3-1)

For example, consider we have two strings "*Table 8.1–Secured housing finance commitments to individuals, ANNUAL*" and "*Table 8.2–Housing finance commitments to individuals, MONTHLY–Seasonally adjusted*". Let minimum words count γ = 3. Let "!@#$%^&*()_+-={}|\:""""?¿/.,<>'¡º×÷';«»[] " be the set of separators. Then the String-sets ΣSs_n are created as shown below.

$$\left\{\begin{matrix} Table\ 8.1\ Secured \\ Secured\ housing\ finance \\ finance\ commitments\ to \\ to\ Individuals\ ANNUAL \end{matrix}\right\} \approx \left\{\begin{matrix} Table\ 8.2\ Housing \\ Housing\ finance\ commitments \\ Commitments\ to\ individuals \\ Individuals\ MONTHLY \end{matrix}\right\}$$

(3-2)

3.3.2. N-gram variation

N-gram variation is the variation in the number of characters that the application will advance when generating each gram. Let us assume that $ngram(\ Ss_1, i, n)$ is the n-gram substring of string S_1 starting at i^{th} position of length n, then the next n-gram substring in the sequence order is $ngram(\ Ss_1, i +\ \nu\ , n)$ where ν is the number of positions (variation) to be moved in addition to the given n-gram sequence order i.e. the number of forward character position advancements to be performed when moving to the next n-gram substring.

For example, consider two strings "*The University of Western Australia*" and *"University of Notre Dame Australia*". Forming 4-gram substring sets with variation ν=1 will create the following gram sets:

$$\{\text{The, univ, iver, ersi, sity, of, west, ster, aust, stra, rali}\}$$
$$\{\text{univ, iver, ersi, sity, of, notr, dame, aust, stra, rali}\}$$

(3-3)

3.3.3. Error base

Error base Ē is the optional percentage of error threshold that is to be tolerated when considering gram matches. Assume that Sg_1 is the n-gram substring of string S_1 and Sg_2 is the n-gram substring of string S_2, Ē is the acceptable margin of percentage error for performing string similarity as shown below.

$$Sg_1 \sim (\, Sg_2 \pm \bar{\text{E}}) \mid \bar{\text{E}} = \frac{|\, Sg_1 - \ Sg_2|}{Sg_1} \times 100$$

(3-4)

For example, assume that the error threshold for gram match is set to 3% and the hash code in consideration is -122 (for gram “The”). Then the error threshold to be tolerated for string similarity is between the upper limit and lower limit which is ± -3.66 (i.e. 3% of -122).

3.3.4. Pair-wise Dissimilarity Matching

The dissimilarity calculations are performed between hash grams of any given two strings, or two strings sets or even between two sets of string sets in order to calculate dissimilarity ratios between hash grams. Consider two strings, S_1 *and* S_2, whose constituent partial hash codes are stored in the hash bags $H_1\left(h_{1,}h_2, h_{3,} \ldots . h_i\right)$ *and* $H_2\left(h_{1,}h_2, h_{3,} \ldots . h_j\right)$. If $l|H_1, H_1|$ is the largest array set count then, the dissimilarity $\mathbb{D}(S_1, S_2)$ between the sets is calculated as shown in equation (3-5). The intuition is that the smaller the h-gram dissimilarity distance, the greater the similarity between two semantic entities.

$$\mathbb{D}\left(H_1, H_{2,}\right) = \frac{\sum_{i=1}^{l|H_1,H_2\,|}\left(100 \cdot \frac{\left|H_1\,(h_i) - H_2\,(h_j)\right|}{Max\left[H_1\,(h_i), H_2\,(h_j)\right]}\right)}{l|H_1, H_2\,|}$$

(3-5)

In order to keep our system more flexible, we have also included conditional pair-wise matching option. For strings having String-sets, the h-gram match can perform incremental pair-wise comparison (if opted) as illustrated in Figure 3.2 up to level σ. The intuition is to support partial string comparison and preservation of strings ontology [2, 78] as part of the matching process.

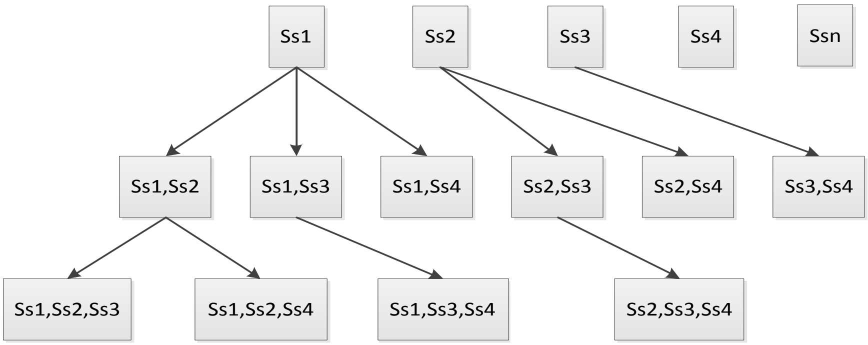


Figure 3.2: Incremental Pair-Wise Comparison up to σ -level Count (default σ = 3, $\Sigma Ss_n \mid n = 4$, configuration value)

Algorithm 1: HGramMatch ($\mathbf{H_{1,}H_2}$)

h-gram record matching algorithm

```
Input:
$\boldsymbol{H_1}$, Set of hash values of entity $\boldsymbol{E_1}$, {$\boldsymbol{h_{1,}h_2, h_{3,}} \ldots . \boldsymbol{h_i}$}
$\boldsymbol{H_2}$, Set of hash values of entity $\boldsymbol{E_2}$, {$\boldsymbol{h_{1,}h_2, h_{3,}} \ldots . \boldsymbol{h_j}$}
Output:
Return sets dissimilarity ratio δ ($\boldsymbol{H_1}, \boldsymbol{H_2}$)
Let ŋ = n-gram variation (see section 3.3.2)
Let Ē = error base threshold (see section 3.3.3)
Let δ = dissimilarity ratio (see section 3.3.4)
// perform one-to-one sequential gram match without error threshold
DO until next partial hash value $\boldsymbol{h_i}$ in $\boldsymbol{H_1}$ = ∅ OR $\boldsymbol{h_j}$in $\boldsymbol{H_2}$ = ∅
    Evaluate ŋ($\boldsymbol{h_{i,}h_j}$):=($\boldsymbol{h_i} \sim \boldsymbol{h_j}$ $\boldsymbol{iff}$ $\boldsymbol{h_i = h_j}$)
SET dissimilarity ratio δ ($\boldsymbol{H_1}, \boldsymbol{H_2}$)
IF δ = 100% do
    SET $\boldsymbol{\Sigma Ss_n}$ = String-sets (see section 3.3.1)
    FOR each String-set $\boldsymbol{\Sigma Ss_n}$ DO
        // perform extended n-gram match with acceptable error threshold
        FOR each partial hash value $\boldsymbol{h_i}$ in $\boldsymbol{H_1}$ = ∅ AND $\boldsymbol{h_j}$ in $\boldsymbol{H_2}$ = ∅ DO
            Evaluate ŋ ($\boldsymbol{h_{i,}h_j}$):=($\boldsymbol{h_i} \sim \boldsymbol{h_j} \pm$ Ē $\boldsymbol{of}$ $\boldsymbol{h_j}$)
        Calculate δ:= δ + $\mathbb{D}(\boldsymbol{H_1}, \boldsymbol{H_{2,}})$
END IF
RETURN δ
```

3.4. Hash gram Record Matching

In order to reduce the number of comparisons required for string comparisons, h-gram algorithm performs record matching in two stages as shown in Algorithm 1. The Algorithm begins with a set of configurable parameters. Let ŋ be the n-gram variation (line 1), Ē be the error base threshold (line 2), and δ be the dissimilarity ratio (line 3). The algorithm begins with a one-to-one sequential gram match (step 4) in which the hash grams are compared in order to perform quick matches between records having minor errors such as typographical constraints, and formatting inconsistencies (step 5-6). Our experimental results have shown that with this technique, strings with slight variations are handled quickly. This approach is also useful to trace short hand notations such as “Jack” for “Jack Smith”, “Mike” for “Michael” etc. as shown in Figure 3.3. Another advantage of having this step is that it significantly reduces the time required to compare possible attribute values such as “Customer Name” column with a “Patient Name” column etc.

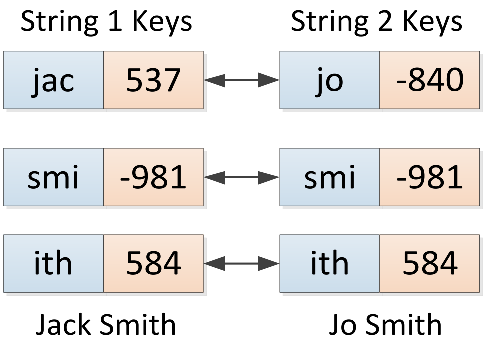


Figure 3.3: One-by-One sequential gram comparison for “Jack Smith” vs. “Jo Smith”

In case the system finds 100% dissimilarity between hash sets (step 7), the system calculates dissimilarity between sets of hash grams (step 8-9). The evaluations are made between hash grams of each of these string subsets (step 10). The accumulated dissimilarity is calculated and evaluated against acceptable dissimilarity threshold (step 11-14). Figure 3.4 shows an example of performing extended n-gram based comparison between strings “The University of Western Australia” and “University of Notre Dame Australia”. In the next chapter, we will describe in detail our proposed framework that allows discovering meaningful data relationships using h-gram record matching as the basis for clustering and cluster mapping.

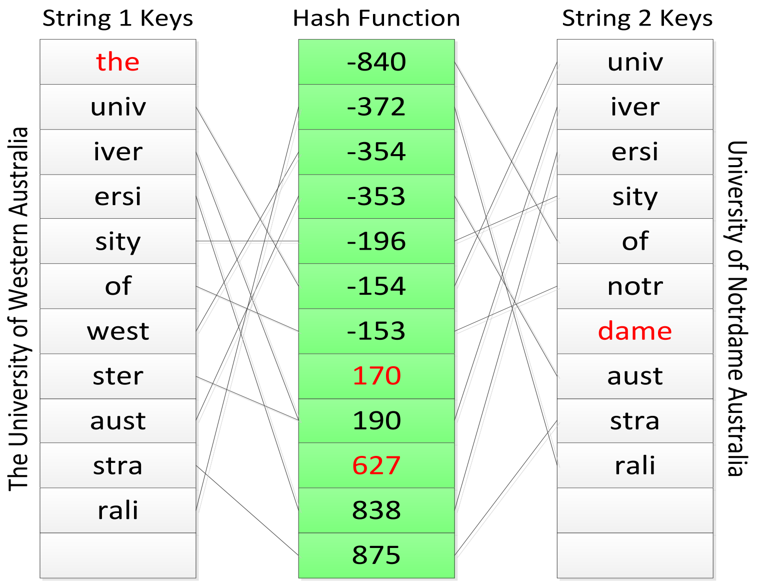


Figure 3.4: Extended n-gram comparison (of hash codes) for “The University of Western Australia” vs. “University of Notre Dame Australia”

3.5. Experimental Evaluation

Our experiments have been carried on Windows Server 2008 server box (Intel Pentium Dual-Core, CPU 3.4 GHz and 2 GB RAM). The prototype is built in C# in Microsoft's Visual Studio 2008 with Microsoft SQL Server 2008 as backend database as visualized in Figure 3.1. Our experimental results were monitored by validating accuracy and performance tests as the benchmark of the evaluation process. Accuracy tests were carried to ensure results obtained are closely associated to expected values. The performance is evaluated as the throughput gained against time and resources. Tests were also carried to assess the workload required, system results throughput and the ability to handle varied sets of data.

Table 3.2: Hash Gram experimental data setup

Database #	Tables	Columns	Rows
The World Bank Data Catalog [24]	4512	67680	2.0 M
The US Federal Govt. Data Catalog [30]	3155	43339	1.4 M
The World Wildlife Fund Data Catalog [33]	21	472	55 K
The Adventure Works Database [42]	254	2608	24 K
National Climatic Data Center [49]	255	1350	15 K
Queensland Govt. Wildlife & Ecosystems [50]	102	259	1.2 K
Medical Data Sets [56]	129	390	33 K

3.5.1. Data Setup

The heterogeneous data [24, 30, 33, 42, 49, 50, and 56] has been collected from different organizations having different sets of data on different domain knowledge. The data used to conduct our experiments are shown in Table 3.2. The total size of the raw data collected is 3.2 GB and the data obtained are in different formats including CSV files, .DAT files, Oracle, and SQL Server databases. The prototype facilitates a number of parameters that the user can setup at run time. Table 3.3 lists these configuration parameters with their default values for h-gram matching technique.

Performance comparison of h-gram technique has been made against closely related algorithms (as analysed in Chapter 2). It is important to note that comparing with all the data linkage methods [18, 34, 37, and 69] is beyond the scope of this research. The relevant techniques used for comparison against our proposed hashing technique are as follows:

1. *Edit distance* (Levenshtein) and Hamming distance were chosen for their wide applicability and popularity above other distance techniques.
2. *Soundex* was chosen for its broad use by various organizations and for its use of phonetics.
3. *n-grams* (for n from 2 to 4) are chosen for its close relation to our proposed h-grams as a base comparison. Two to four were chosen as parameters because previous research [18, 34] shows that most practical results fall between 2 and 4.
4. *IR-Sum* of Set-of-sets [8] was chosen for its close relation to our proposed h-gram function as an enhancement of basic n-grams.
5. *Synonyms-based* or dictionary-based comparison was chosen for highlighting the quality of the sample data, showing the strictness of the applied standards.

The comparison against aforementioned methods demonstrate an adequate variety of data linkage techniques as the above methods represent commonly used, proven methods in a range of areas [34, 37] as well as methods closely relating to our proposed h-gram record matching process.

Table 3.3: Configuration Parameters

Parameter	**Range**	**Default**
n-gram	1 to 5	3
n-gram variation	1 to 5	1
hash gram digit	1 to MAX	4
String set	1 to 5	1
error %	0 to 10%	5%

3.5.2. Accuracy Metrics

Record Matching problems don’t arise from one type of error alone. Therefore, any experiment that is to be considered accurate would need to include an accurate representation of scenarios that consider all common types of record matching problems. Only when there is an accurate representation of error types, can a reliable comparison against other techniques be made. To solve the problem of performing a reliable comparison against the selected methodologies, we performed accuracy tests across different domain table attributes accounting for a variety of possible scenarios (phonetic similarity, abbreviations, mistypings, etc.). Our manual selection of data samples ensured that multiple error types were evenly represented in the experiment. However, another problem which must be noted is that all the methods selected for comparison are based on absolute similarity scales. Thus, in order to perform the intended comparison, it was necessary to create a system to relativise the methods to be compared against our proposed h-gram technique through a set of similarity rules.

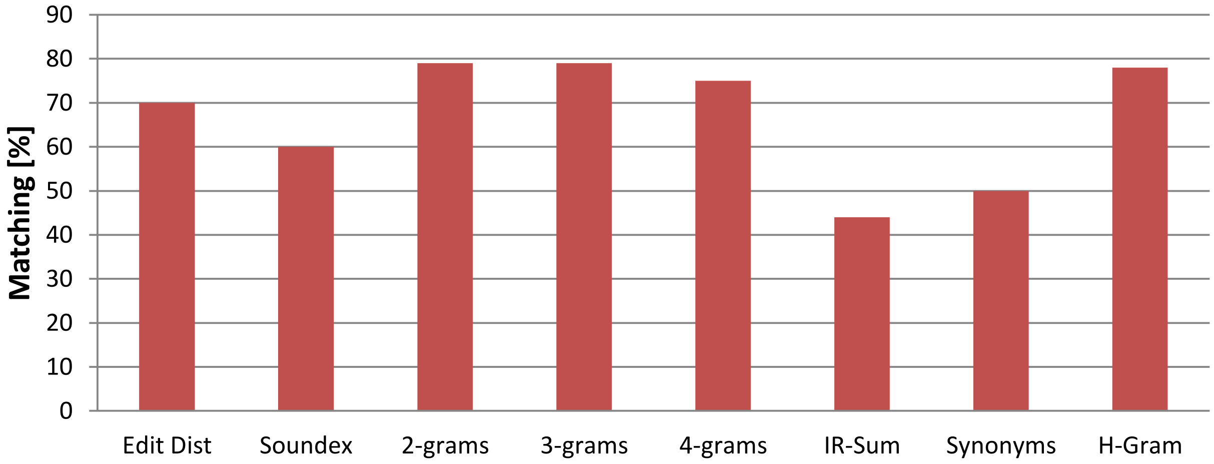


Figure 3.5: Matching percentages by applying each method to the sample data.

Figure 3.5 illustrates a comparison of matching percentages from the methods under consideration. The displayed percentages are referred to the dataset utilized to experimentally validate our method. The main reason for including the synonym (dictionary) based approach was to highlight the quality of the sample data by demonstrating the strictness of the applied standards. On the other hand, measuring the performance is substantially more difficult for the dictionary-based approach than for any of the others and the applicability of the conclusions would not be as solid as the following: 1) CPU/memory usage and 2) Time requirements strongly depend on the specific conditions under which the synonyms are gathered (i.e. over the internet, via intranet from an external database, or from another table in the same database) as well as the size/level of detail (number of synonyms) of the given repository.

Of the experimented methods, IR-Sum [8] of Set-of-sets is the only method which may be considered outside of the string similarity metrics and is therefore most similar to our proposed approach for the reason that both IR-Sum and our proposed method rely on two identical steps: hashing (conversion of stings into numeric facts) and hash code comparison. Based on the matching percentages as shown in Figure 3.5, the accuracy of IR-Sum is significantly lower than that of our proposed method due to the fact that IR-Sum fails to maximize the full potential of the generated hash codes performing detailed-enough comparisons among all the constituents (tuples/grams), but by utilizing a pseudo-random approach (accounting only for the smallest hash value) in order to reduce computational requirements. On the other hand, this aim has been reached through our h-gram method.

Within the string similarity methods, the Synonym method shows the worst accuracy. Distance method (Levenshtein and Hamming) accuracy somewhat better, although it is still too low. As expected, the n-gram based approaches show the highest matching percentage because they represent the most detailed analysis of string comparison: comparing as many combinations of constituent elements as allowed (by the gram type, n, under consideration). It is not clear which sub-type (2-tuples, 3-tuples or 4-tuples) would provide the best performance independently upon the input conditions (strings of substantial length variance such as complete sentences as opposed to single words). Nevertheless, the accuracy of the three tuple methods should be similar for the most common applications. The accuracy of our proposed h-gram method is equivalent to the n-grams method. Identical accuracy is explained by the fact that, at the comparison level, its accuracy is (worst case scenario) as good as that of a 3-gram-based approach: the gram-by-gram might find a match/mismatch which the 3-grams cannot find.

3.5.3. Performance Metrics

The analyzed methodologies might be implemented in various ways, not only in regards to the algorithm, but also with hardware reliance (exclusively depending upon local memory, writing to/reading from temporal files, relying on a local DB, etc.). Thus, accurately comparing methodology performance could prove to be rather complicated. Nevertheless, we have reduced uncertainty in the analysis as much as possible by:

- Creating algorithms from scratch for the selected methodologies (see section 3.5.1), which follow equivalent programming guidelines, and
- Ensuring simulation conditions (same computer[1] under same workload, performance retrieving points at equivalent "positions", etc.) are as similar as possible from one case to the next.

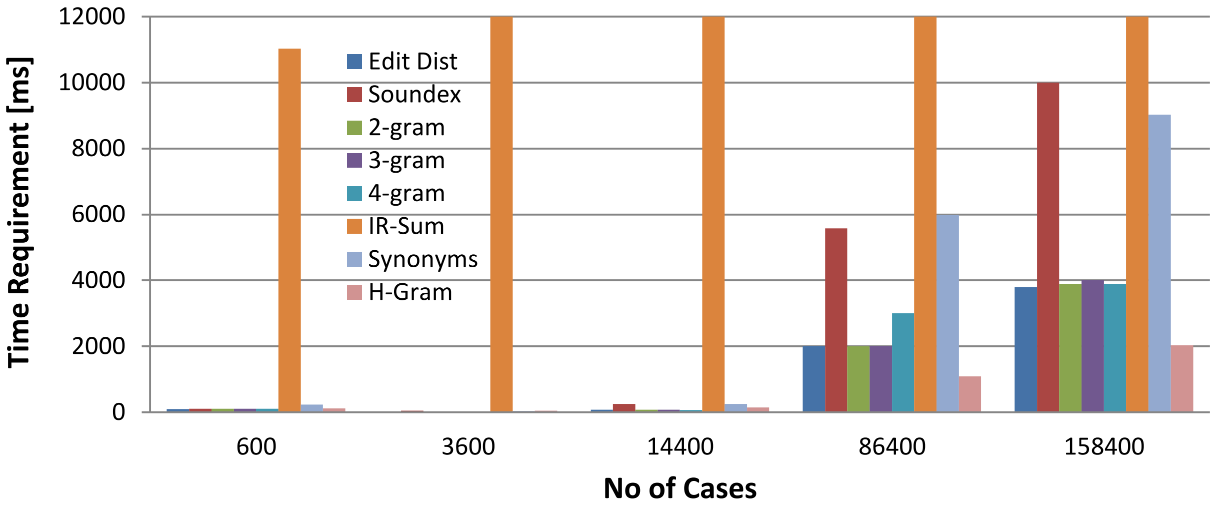


Figure 3.6: Time requirements of the studied techniques between different methods and under different data source sizes

We have measured variables accurately describing the performance from each methodology. Due to peculiarities, comparisons are included in two different groups:

First group of comparisons accounting for variable-size inputs: the specific variable (time requirements) does not show a similar evolution for all methodologies. Time invested by each algorithm to perform all the calculations for the specific data set input as shown in Figure 3.6.

[1] Windows Server 2008, Pentium Dual-Core E5200 @ 3.4GHz

From the experimental results shown in Figure 3.6, we can see h-gram method performs with the least amount of time while edit distance and n-gram based methods follow a more or less similar evolution. The large difference between IR-SUM and other methods provides a good idea about the main drawback from the hashing methods: hashing-un-hashing, at least by using conventional approaches, is too computationally expensive. IR-SUM requires conversion of all the data to find the smallest hash value which are considered for evaluation. The reason why our method does not show a so big difference with respect to the string-comparison methods is because of the small size of our "hash gram codes". Instead of having a huge hash bag accounting for any eventual entry, which has to be inspected every time that the hashing/un-hashing process occurs, we are relying on a simple array with small dimensions, whose access/retrieve times are much smaller than the ones from a conventional hash table.

Second group of comparisons performed for a single input: variations in the size of the input set do not affect comparisons; evolutions of the different methodologies remain virtually identical independent of the size of the input. Both variables are referred to average values of intermediate tracking points (located in equivalent positions throughout each algorithm). Variables included in this group are:

- Average values for percentage of CPU usage and
- Average values for RAM memory usage (virtual MB)

The experimental results for CPU and RAM usages are displayed in Figure 3.7 and Figure 3.8. We can see h-gram uses reasonably less system resource when compared with other methods excepting 3 and 4-grams. Nevertheless, run-time cost of h-gram method is much lower when compared with 3 and 4 gram results (see Figure 3.6).

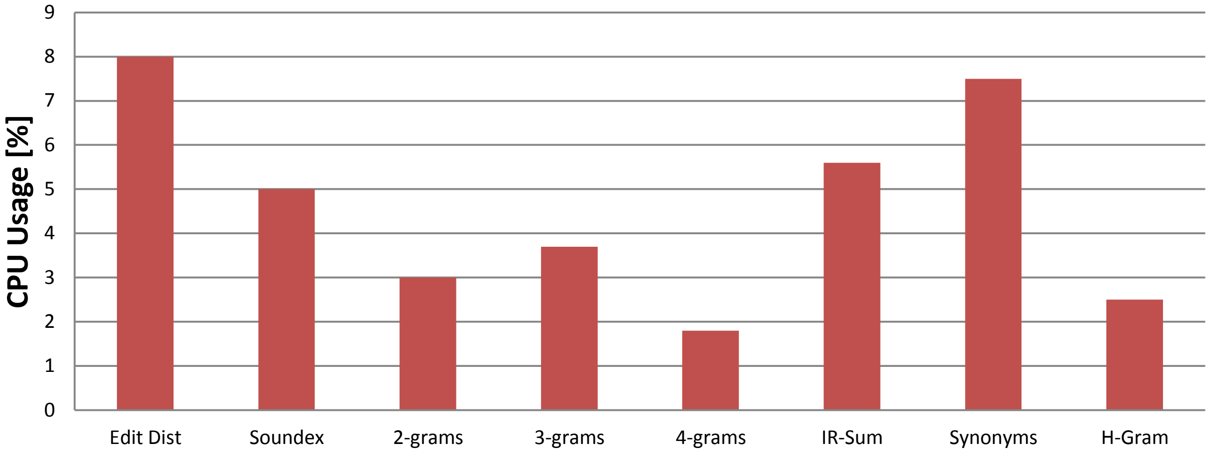


Figure 3.7: Average CPU usage

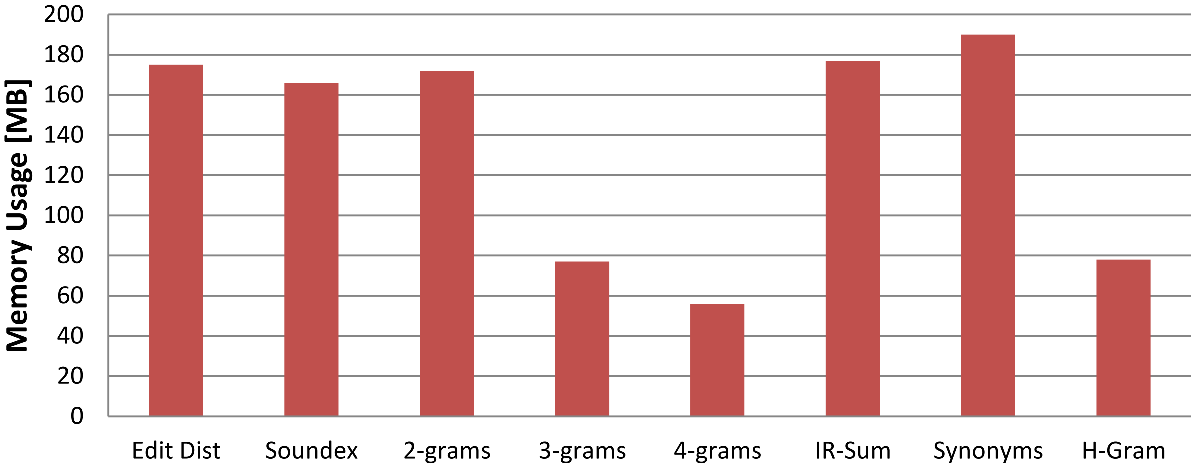


Figure 3.8: Average Memory usage

3.6. Summary

In this chapter, we examined how to quickly perform record matching using our newly proposed h-gram matching technique. We showed how h-gram technique helps organizations in performing various data intrinsic tasks including Data Linkage, Record Matching, Data Cleaning, Data Migration, and Semantic/Faceted browsing. In summary, we showed how our approach can help different anticipating bodies towards probabilistic matching and discovering required matching data during data integration and other data intrinsic tasks without prior knowledge of the data files, having little or no documentation and without waiting for a long delay at run time. Through experimental results, we prove that our technique performs superior record matching with h-gram technique when conducting linkages with data sets containing up to several hundred thousand records.

In the next chapter, we will show how h-gram technique is used to build clusters of related taxonomy definitions with small dissimilarity distances in our newly proposed Ontology Guided Data Linkage (OGDL) framework. We will also show how our approach can provide a good balance of accuracy vs. computational requirements and can significantly reduce cluster sizes at multiple levels. Through this approach, we also prove that similar entities of one cluster are dissimilar to entities of other clusters.

Chapter 4
Ontology Guided Data Linkage (OGDL) Architecture

4.1. Problem Description

In this chapter, we consider the problem of constructing a ‘knowledge based’ multi-faceted cluster mapping technique, which aims at extracting probable relationships between correlated data clusters on a structure level. We formally introduce the linkage problem through our Ontology Guided Data Linkage (OGDL) framework and show how our algorithms can be applied to variable databases. The proposed framework intends to create a feasible method for discovering related information, as part of bottom-up system managed process that allows top-down information extraction procedures using user-friendly queries. Our main methodology intuition is that end-users performing semantic queries will not have knowledge of meaningful data relationships unless relationships have been established, and related information presented. Figure 4.1 shows the architecture of our framework. When the user creates a query or intends to visualize relationships on a graph, the results delivered include not only the data requested by the user, but also includes directly related data. The experiments illustrated in this chapters show that our approach to the consideration of the relative importance of ontological information in input data shows promising findings. To the best of our knowledge, this is the first attempt initiated to solve data linkage problem using a multi-faceted cluster mapping strategy, and we believe that our approach presents a significant advancement towards accurate query answering and future real-time online semantic reasoning capacity.

Although the proposed approach is designed primarily for the acquisition of data linkage implementation intelligence, it is applicable to a variety of data discovery purposes. The application can be used by data managers, researchers, or analysts, for a variety of purposes, including optimized multi-domain knowledge representation, as it doesn’t require a data structure or complex query knowledge. The series of steps performed as part of the OGDL framework is aimed at discovering data linkages between large-size databases, with minimal user involvement. In other words, in this research, the proposed framework directly addresses high computational overheads through a multi-layer strategy which significantly reduces the amount of data considered for comparison at subsequent stages, and which enables cluster self-expansion through the construction of ontology guided data linkage structures.

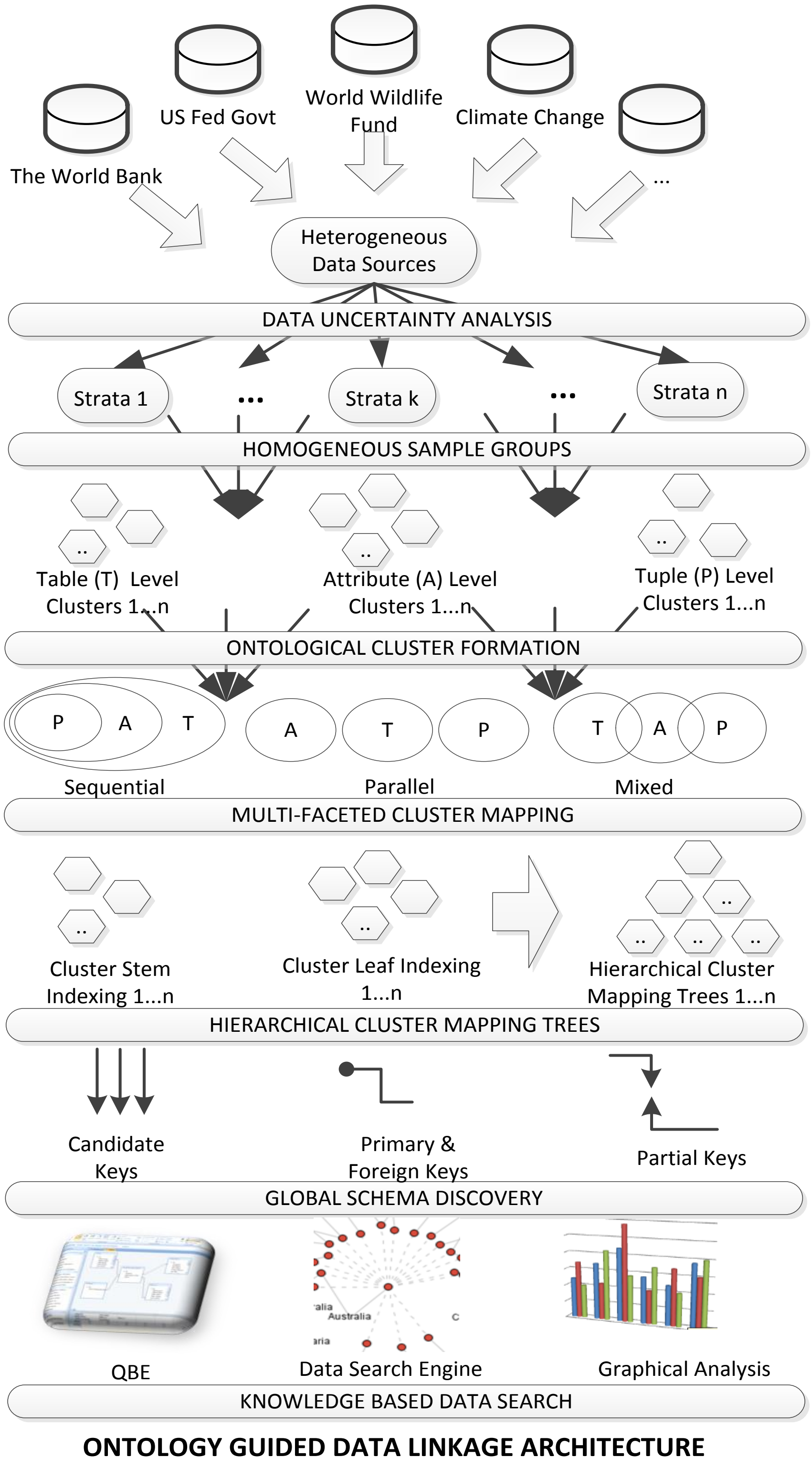
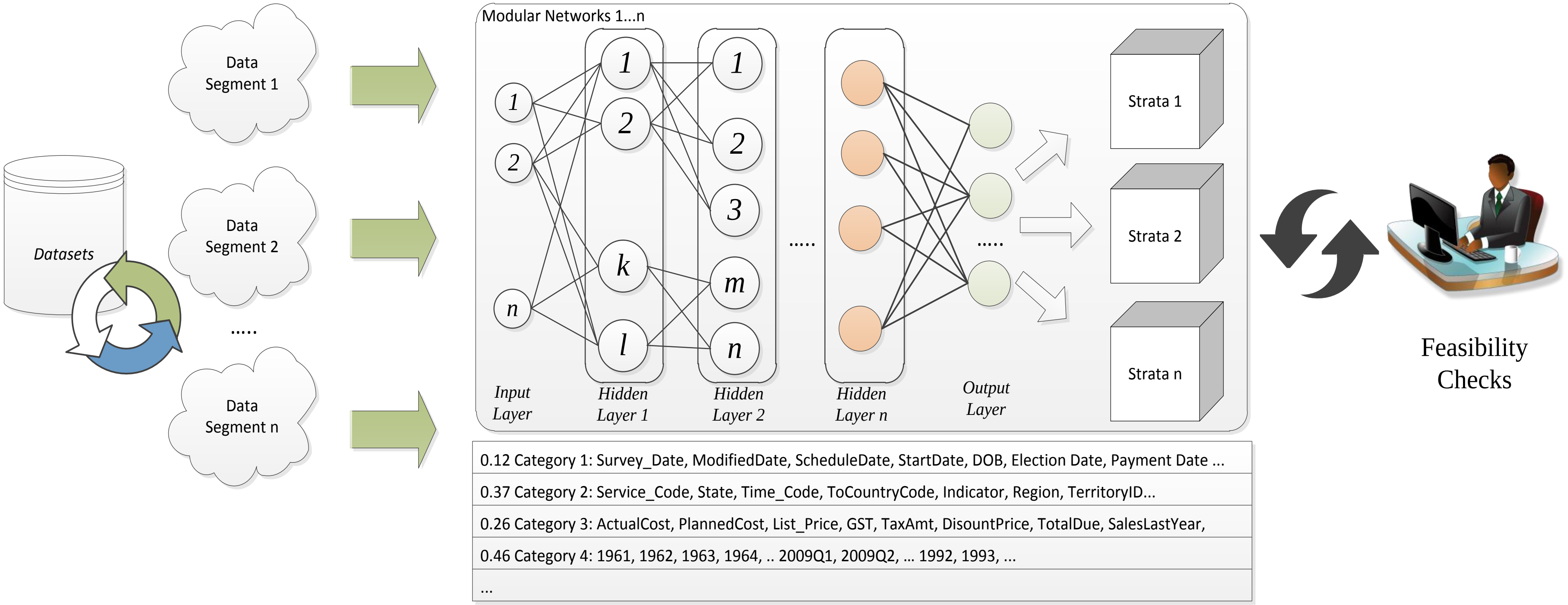

US Fed Govt
World Wildlife Fund
Climate Change
The World Bank
...
Heterogeneous Data Sources
DATA UNCERTAINTY ANALYSIS
Strata 1
...
Strata k
...
Strata n
HOMOGENEOUS SAMPLE GROUPS
..
Table (T) Level Clusters 1...n
Attribute (A) Level Clusters 1...n
Tuple (P) Level Clusters 1...n
ONTOLOGICAL CLUSTER FORMATION
P
A
T
A
T
P
T
A
P
Sequential
Parallel
Mixed
MULTI-FACETED CLUSTER MAPPING
Cluster Stem Indexing 1...n
Cluster Leaf Indexing 1...n
Hierarchical Cluster Mapping Trees 1...n
HIERARCHICAL CLUSTER MAPPING TREES
Candidate Keys
Primary & Foreign Keys
Partial Keys
GLOBAL SCHEMA DISCOVERY
Australia
QBE
Data Search Engine
Graphical Analysis
KNOWLEDGE BASED DATA SEARCH
ONTOLOGY GUIDED DATA LINKAGE ARCHITECTURE

Figure 4.1: The general architecture of OGDL Framework

A key component of our framework is a novel faceted search engine for visualizing mapping clusters (see Figure 4.12Figure 4.11). In the experimental section, we show how cluster mapping trees act as concept graphs that can support well-informed accountable governance decisions, as machine learning recommendations to human experts. We show that our search engine runs in polynomial time and that it is highly effective in a real-world scenario. Specifically, this research addresses the semantic reasoning and data integration problem by providing an intelligent multi-layer cluster formation and multi-faceted cluster mapping approach to integrate and easily analyse multi-domain information. We propose a paradigm where the user can interact, using user-friendly queries, with ontological structures by searching for a few key words. Bottom-up input data extraction of semantic knowledge is enabled by system configurations on connections between semantically equivalent tables, attributes and tuples.

4.2. Data Uncertainty Analysis

When dealing with large volumes of data (numeric; categorical; string based; etc.) obtained from different sources, we are vulnerable to different types of ‘data uncertainties’ such as different formats; Null values; length constraints; typographical errors; and shorthand notations, which may well be one of the biggest obstacles to performing successful data linkage. An important initial step for successful linkage is data cleaning and standardization, as noisy, incomplete and incorrect information is common in real-world databases [7]. In an effort to improve the quality of such data, we employ preliminary ‘data uncertainties’ process steps to ensure optimal results. An important advantage of the ‘data uncertainties’ process is that it is a ‘one-off’ system handle process aimed at cleaning, classifying and organizing observational data, to minimize any manual effort. It has to be noted that these steps are aimed at approximating the quality of data and doesn’t guarantee that it overcomes all data uncertainties.



0.12 Category 1: Survey_Date, ModifiedDate, ScheduleDate, StartDate, DOB, Election Date, Payment Date ...

0.37 Category 2: Service_Code, State, Time_Code, ToCountryCode, Indicator, Region, TerritoryID...

0.26 Category 3: ActualCost, PlannedCost, List_Price, GST, TaxAmt, DisountPrice, TotalDue, SalesLastYear,

0.46 Category 4: 1961, 1962, 1963, 1964, .. 2009Q1, 2009Q2, ... 1992, 1993, ...

...

Figure 4.2: The design and evolution of a Multi-Modular Neural Networks architecture aimed at data classification

4.2.1. Multi-Modular Network Classification

Ontology based classification is the process of categorizing given data into predefined classes, based on similarities in their ontologies. The Neural Network approach [2] attempts to introduce intelligent behaviour by clustering attributes into categories using attribute properties. However, Neural Networks are highly expensive when applied to datasets with thousands of table columns. This is due to the fact that attribute similarity comparisons increase at an increasing rate as nodes are continuously added to the input layer. The computation time is dependent on these node counts and their correlations, and thus the increasing number of nodes has a direct impact on its performance. In order to reduce the number of comparisons, we introduced an extended version of Modular Neural Networks [81], using the process of ranking and classifying ontological characteristics in multiple modules. Figure 4.2 shows the design and evolution of a Multi-Modular Neural Networks architecture aimed at creating and classifying different strata samples with particular classified attributes. Consider The World Bank database [24] as the data segment inputs into the classification process. Then the resulting output is the classified attribute categories, such as category of dates {Survey_Date, ModifiedDate, ScheduleDate etc.} as found when applying the architecture against The World Bank database [24]. Some of the examples of these categories are listed in Figure 4.2 Table.

The key advantages of the modular networks approach include: reduction of hidden layer complexity; support for data fusion and average prediction making; combination of multiple techniques; concurrent execution of multiple network models with high robustness; and fault-tolerant average results [82]. We generate multiple modular networks in parallel, working independently on different data segments. Therefore the speed of the system is nearly independent of the number of network layers involved in the classification process. We used different hidden layers (modules) that can inter-connect with modules of other types. At a very basic level, hidden layers are the building blocks for OGDL ontological data-modelling. Each hidden layer level is recognized as a subset of input data. The hidden layers are considered in a hierarchical order, where the output of the first layer is given as an input to the next hidden layer. Parallel neural network architecture is superior to a single neural network, as proposed in [82]. At a higher level, stratified sample sets are constructed on top of neural network classified attributes. Our proposed approach to modular classification networks can be constructed on structures with greater complexity.

We employed the Spearman's Rank correlation coefficient [58] to determine the relationship strength between attributes associated with hidden layer tasks, based on its predicted ranking weight, as an average of rankings made by the participating judges. The results provided robust similarity measures as well as sound reliability measures. Such statistical guarantees have been an existing challenge in existing classification solutions. In our approach, the *Judges* are pattern matching agents, for example for the recognition of *m* number of pattern clusters for the given *n* number of features. The multi-modular network measurement approach ranks the ontological similarity of the input layer variables in different layers, using classified definitions available within the hidden layer. The output layer demonstrates homogeneous groups of ontologically matched attributes, organized in categories.

The ontological classification of attributes is optimized as shown in equation (4-1). Assume that a group of *n* entity values are arranged in an order of merit with tics $c_1, c_2, \dots c_n$. Let $\rho(c_i, c_j)$ be the ontological similarity rank defined between two characteristics c_i and c_j. A pair of attributes is determined by the similarity of its features and is defined as the maximum $\rho\ (c_i, c_j)$. The averages of such classified attributes are calculated based on their related best correlation weights, as obtained by parallel modular network results, with the intuition that this would best approximate an accurate classification solution, and produce results quickly because of the use of concurrent computing.

$$\rho\ (c_i, c_j) = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_i^2} = 1 - \frac{6\ \sum_{i=1}^{n} d_i^2}{n\ (n^2-1)} \mid \sum_i d_i^2 = \sum_i [(x_i - \bar{x}) - \ (y_i - \bar{y})]^2 \tag{4-1}$$

4.2.2. Stratified Sampling

In the case where there are an unmanageable number of records, the computational requirements become unacceptably high, thus necessitating sampling methodologies. Sampling is the statistical practice concerned with the selection of an unbiased or random subset of individual observations, within a population (dataset) of individuals [58]. In our previous step, we already classified attributes into various modules based on their ontological characteristics. Hence, we need to choose a sampling technique that can effectively reduce the sample population heterogeneity, provide more representatives and greater accuracy. Hence, we employed stratified sampling process which supports to extract knowledge from relatively large volumes of data and supports making predictions based on statistical inferences. In a stratified random sampling process, the data is divided into homogenous groups called strata. Each stratum differs from another stratum, but is homogenous within itself [58]. The advantage of stratified sampling is that it improves precision, which is especially required when working with small samples.

We take advantage of results from the aforementioned multi-modular network classification process in order to create strata samples. The equation for stratified sampling in (4-2) shows how the number of samples variable is calculated for a stratum *i,* in reference to *n,* the total number of strata. The number of samples taken in each stratum is a function $f(i)$ of its standard deviation st_{dev} , that is, the stratum's homogeneity: the more homogeneous the stratum (the smaller its standard deviation); the fewer the number of samples that is taken.

$$f(i) = n \cdot \frac{st_{dev(i)}}{\boldsymbol{Max}\{st_{dev}\}} \quad No.\ Samples(i) = \begin{cases} 0.5 \cdot n, & f(i) > 0.5 \cdot n \\ f(i), & 0.5 \cdot n \geq f(i) \geq 0.1 \cdot n \\ 0.1 \cdot n, & f(i) < 0.1 \cdot n \end{cases} \tag{4-2}$$

4.3. Multi-Layer Ontological Cluster Formation

Clustering is the task of organizing data into groups (clusters) such that similar or close data objects are put in the same cluster [86]. We define ontological clustering as the process of globally and significantly reducing the number of semantic entities at multiple levels, with the aim to reduce its data linkage computational expense. We build clusters in result to hash grams of related taxonomy definitions with small h-gram dissimilarity distances. This approach provides a good balance of accuracy vs. computational requirements and significantly reduces cluster sizes at multiple levels when compared to similar approaches. This approach also ensures that similar entities of one cluster are dissimilar to entities of other clusters. One of the most popular clustering algorithms used in scientific and industrial applications is the k-means clustering algorithm [18] and its derivatives, such as the PAM (Partitioning Around Medoids) algorithm. However, the k-means algorithm requires the number of clusters to be defined in advance; is highly sensitive to outliers; and is not suitable for representing hash grams of variable size. Hence, we designed our clustering process through an agglomerative approach. Our algorithm starts with all the entities forming separate clusters which then merge with ‘close’ clusters until a dense cluster is formed that contains all objects.

Algorithm 2.1: OGDL Multi-Layer Cluster Formation on Tables

OGDL table level cluster formation algorithm

```
Input:
- A set of modular network classifiers: $\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \boldsymbol{\psi_3} \dots \boldsymbol{\psi_n}$
- A set of stratified sample sets: $\boldsymbol{\delta_{1,a_1}}, \boldsymbol{\delta_{2,a_2}}, \boldsymbol{\delta_{3,a_3}} \dots \boldsymbol{\delta_{n,a_n}}$
- A hash gram generation function: h-gram (see section 3.3.2)
- A hash gram dissimilarity function: $\mathbb{D}$ (see section 3.3.4)
- confidence level: $\rho$ // default .75 value set by the user at runtime
- Number of nearest ontological neighbours to consider: $k$

Output:
A set of table level clusters $\Sigma\boldsymbol{C_{j,k}} \mid \boldsymbol{j} := \mathbf{1,2,3} \dots \boldsymbol{n\ and\ k} \coloneqq \{\boldsymbol{\tau}\}\ \boldsymbol{\tau} := \textbf{table level}$
Initialise empty cluster $\Sigma\boldsymbol{C_{j,k}} \coloneqq \{\varnothing\}$
FOR $\forall\ \boldsymbol{t} \in \mathbb{T}$: // table level clusters
    $\boldsymbol{\varphi_t}$ := h-gram($\boldsymbol{t}$)
    $\boldsymbol{\varphi_t}' \equiv \forall$ (k+1) h-gram(Ō[$\boldsymbol{t}$]) weights [][]→ $\boldsymbol{\varphi_t}$
    WHILE $\Sigma\boldsymbol{C_{j,k=\tau}} \neq \varnothing$:
        IF $\mathbb{D}(\boldsymbol{\varphi_t}, \boldsymbol{C_j}) \sqsubseteq \rho$ OR $\mathbb{D}(\boldsymbol{\varphi_t}', \boldsymbol{C_j}') \sqsubseteq \rho$ |
        $\boldsymbol{C_{j,k=\tau}}' \equiv \forall$ (k+1) h-gram(Ō[$\boldsymbol{C_{j,k=\tau}}$]) weights [][] → $\boldsymbol{C_{j,k=\tau}}$
            $\Sigma\boldsymbol{C_{j,k=\tau}} \coloneqq \Sigma\boldsymbol{C_{j,k=\tau}} \cup \boldsymbol{t}$ // merge with existing cluster
        ELSE
            $\Sigma\boldsymbol{C_{j,k=\tau}} \mathrel{+}= \Sigma\boldsymbol{C_{j,k=\tau}}(\boldsymbol{t})$ // form a new cluster
        END IF
NEXT

RETURN $\Sigma\boldsymbol{C_{j,k}}$
```

Through our experiments, (see section 4.9), we prove that this agglomerative clustering approach represents an effective way to achieve probabilistic matching with both high accuracy and acceptable computational cost, when compared to similar approaches. In order to discover the flow of semantic information in multiple dimensions, the framework performs a multi-level clustering process at table, attribute and tuple levels, as shown in Figure 4.3 (b), with the aim to capture different layers of ontologies and their interrelationships. As can be seen from the diagram, the results obtained from table level clusters are used to perform structural level matching; those from attribute level clusters are used to perform schema level matching; and the results obtained from tuple level clusters are used to perform key word scans and hierarchical data matching.

Algorithm 2.2: OGDL Multi-Layer Cluster Formation on Attributes and Tuples

OGDL attribute level and tuple level cluster formation algorithm

Input:
- A set of modular network classifiers: $\boldsymbol{\psi_1, \psi_2, \psi_3} \dots \boldsymbol{\psi_n}$
- A set of stratified sample sets: $\boldsymbol{\delta_{1,a_1}, \delta_{2,a_2}, \delta_{3,a_3}} \dots \boldsymbol{\delta_{n,a_n}}$
- A set of table level clusters: $\Sigma\boldsymbol{C_{j,k}} \mid \boldsymbol{k} \coloneqq \{\boldsymbol{\tau}\}$
- A hash gram generation function: *h-gram* (see section 3.3)
- A hash gram dissimilarity function: $\mathbb{D}$ (see section 3.4)
- confidence level: ρ // default .75 value set by the user at runtime
- Number of nearest ontological neighbours to consider: *k*

Output:
A set of complete *OGDL clusters* $\Sigma\boldsymbol{C_{j,k}} \mid \boldsymbol{j} \coloneqq \mathbf{1, 2, 3} \dots \boldsymbol{n}$ $\boldsymbol{and}$ $\boldsymbol{k} \coloneqq \{\boldsymbol{\tau, \partial, \rho}\}$
$\boldsymbol{\tau}\text{: = }\textbf{table level}, \boldsymbol{\partial}\text{: = }\textbf{attribute level}\ \boldsymbol{and}\ \boldsymbol{\rho}\text{: }\textbf{tuple level}$

```
WHILE $\boldsymbol{\psi_n} \neq \varnothing$:
  FOR ∀ $\boldsymbol{a} \in \boldsymbol{\psi_n}$: // attribute level clusters
    $\boldsymbol{\varphi_\partial}$ := h-gram($\boldsymbol{\partial, a}$)
    $\boldsymbol{\varphi_\partial}'$ ≡ ∀ (k+1) h-gram(Ō[$\boldsymbol{\partial, a}$]) weights[][] → $\boldsymbol{\varphi_\partial}$
    WHILE $\Sigma\boldsymbol{C_{j,k=\partial}} \neq \varnothing$:
      IF $\mathbb{D}(\boldsymbol{\varphi_\partial, C_j}) \sqsubseteq \rho$ OR $\mathbb{D}(\boldsymbol{\varphi_\partial}', \boldsymbol{C_j}') \sqsubseteq \rho$ |
    $\boldsymbol{C_{j,k=\partial}}'$ ≡ ∀ (k+1) h-gram(Ō$[\boldsymbol{C_{j,k=\partial}}]$) weights[][] → $\boldsymbol{C_{j,k=\partial}}$
        $\Sigma\boldsymbol{C_{j,k=\partial}} \coloneqq \Sigma\boldsymbol{C_{j,k=\partial}} \cup \boldsymbol{\partial}$ // merge with existing cluster
      ELSE
        $\Sigma\boldsymbol{C_{j,k=\partial}} \mathrel{+}= \Sigma\boldsymbol{C_{j,k=\partial}}(\boldsymbol{\partial, a})$ // form a new cluster
      END IF
    END WHILE
    FOR ∀ $\boldsymbol{\rho} \in \boldsymbol{\delta_{n,a}}$: // tuple level clusters
      $\boldsymbol{\varphi_\rho}$ := h-gram($\boldsymbol{\rho, a}$)
      $\boldsymbol{\varphi_\rho}'$ ≡ ∀ (k+1) h-gram($\mathbb{T}$[$\boldsymbol{\rho, a}$]) weights[][] → $\boldsymbol{\varphi_\rho}$
      WHILE $\Sigma\boldsymbol{C_{j,k=\rho}} \neq \varnothing$:
        IF $\mathbb{D}(\boldsymbol{\varphi_t, C_j}) \sqsubseteq \rho$ OR $\mathbb{D}(\boldsymbol{\varphi_t}', \boldsymbol{C_j}') \sqsubseteq \rho$ |
      $\boldsymbol{C_{j,k=\rho}}'$ ≡ ∀ (k+1) h-gram($\mathbb{T}[\boldsymbol{C_{j,k=\tau}}]$) weights[][] → $\boldsymbol{C_{j,k=\rho}}$
          $\Sigma\boldsymbol{C_{j,k=\rho}} \coloneqq \Sigma\boldsymbol{C_{j,k=\rho}} \cup \boldsymbol{\rho}$ // merge with existing cluster
        ELSE
          $\Sigma\boldsymbol{C_{j,k=\rho}} \mathrel{+}= \Sigma\boldsymbol{C_{j,k=\rho}}(\boldsymbol{\rho, a})$ // form a new cluster
        END IF
      END WHILE
    NEXT
  NEXT
END WHILE
RETURN $\Sigma\boldsymbol{C_{j,k}}$
```

The algorithm for creating multi-layer clusters at table level is shown in Algorithm 2.1. The algorithm commences by initializing each table level, attribute level and tuple level cluster set (step 1). The clustering algorithm then pairs items that reference the same table level entity (step 2). Once the initial table level clusters are formed, the algorithm continues to increase the density of these clusters (step 4-5) through merging with entities that have similar ontologies. Similarities are calculated using the ‘dissimilarity’ function $\mathbb{D}$, i.e. the degree of difference between the hash grams of two entities, thus determining the similarity for matching (step 6-10). Dissimilarity with a score of 0 is a perfect match, and a score of 1 represents a ‘null’ match.

Once the table level clusters are formed, the framework performs attribute level and tuple level clusters. The algorithm for creating multi-layer clusters at attribute level and tuple level is shown in Algorithm 2.2. The algorithm commences by inputting table level cluster set formed in the previous stage (algorithm 2.1). The clustering algorithm pair’s items that reference the same attribute level entity based on h-gram matching weights (step 1-5). Once the initial clusters are formed, the algorithm continues to increase the density of these clusters through merging with entities that have similar ontologies. Similarities are calculated using the ‘dissimilarity’ function $\mathbb{D}$, i.e. the degree of difference between the hash grams of two entities, thus determining the similarity for matching (step 6-11). The same logic is applied within each attribute level clusters to discover tuple level clusters. The algorithm pair’s tuple items that reference the same entity (step 12-15). Similarities are calculated using the ‘dissimilarity’ function to determine the cluster matches (16-21). Dissimilarity with a score of 0 is a perfect match, and a score of 1 represents a ‘null’ match.

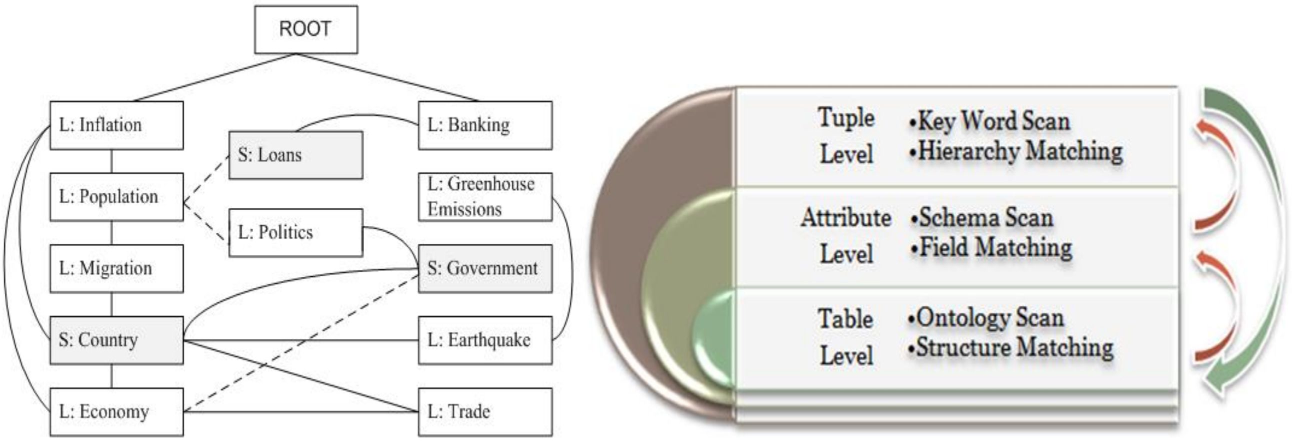


Figure 4.3: (a) The left diagram shows example of clusters with stem (S) and leaf (L) pairs formed in the World Bank Statistical Indicators dataset [23]; (b) The right diagram shows the repetition of cluster formation processes at different levels.

Linguistic taxonomy data inputs are analysed and h-grams are generated for the purpose of similarity matching (step 4). The transformed h-grams are used to make comparisons between data inputs, and the dissimilarity $\mathbb{D}$, as per the accepted convergence level ρ, is determined between different datasets (step 5-6). The proximity results determine whether to merge a cluster with its closest cluster; or to insert it as a new cluster (step 7-10). This approach groups clusters which are semantically similar into larger, more condensed clusters, thus producing a tree of similar concepts. A similar logic is employed iteratively at the attribute level (step 13-18) within each of the classified groups (see section 4), with the resulting grouping either being merged with existing clusters, or added as a new cluster (step 19-21). At the tuple level, a similar logic is applied, except that only strata samples are analysed and that the clusters are developed using frequent pattern mining (step 23-31).

Figure 4.3 (a) shows an example of stem-and-leaf mapping formed in the World Bank Database [10]. The stems {Country, Loans, and Government} are associated with their corresponding leaf structures. As can be seen from the diagram, one of the primary tasks of ontological clustering process is that it reduces the unmanageable number of cluster pairs and maps direct (such as between Country and Trade) and indirect relationships (such as between Population and Politics) which can be used as a basis for matching clusters which is detailed in the next section.

4.4. Multi-Faceted Cluster Mapping

The extraction of meaningful data facts relies not only on the discovery of different sets of ontological clusters: In addition, hierarchical relationships have to be established between clusters. We introduce the ‘multi-faceted’ cluster mapping strategy in order to capture structural relationships between different ontological clusters in different arrangements, which provide us with an advantage to the discovery of hierarchical relationships when compared to existing approaches. This approach is used to arrange the clusters into sequential, parallel and mixed facets, as shown in Figure 4.4. Relative mappings and their relationship strengths are determined and preserved in Attribute-Relation File Format (ARFF) [17].

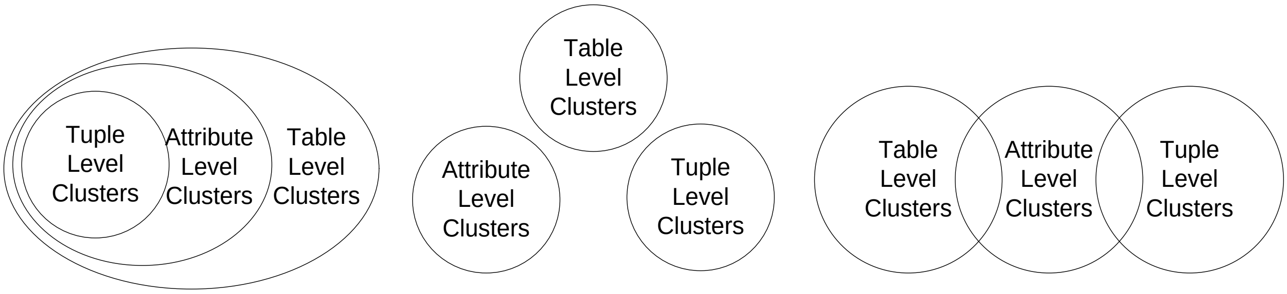


Figure 4.4: Multi-faceted cluster mapping arrangements; (a) sequential cluster mapping; (b) parallel cluster mapping; (c) mixed cluster mapping

In order to rank the correlation between data structures as unordered clusters rather than as paired observations, we used the Intra-class correlation (ICC) technique [58], with the intuition that the ICC is suitable due to the unordered state of the data, in contrast to the Pearson correlation coefficient. A high-level cluster mapping algorithm to determine how strongly units in similar clusters resemble one another is shown in Algorithm 3. If γ $^{\rightarrow}$ and δ $^{\rightarrow}$are the ontological vectors of two cluster entities γ and δ in cluster space, then their similarity cluster relationship is determined through the correlation calculation of Ψ, as shown in the below equation.

$$\Psi(\gamma,\delta) = \frac{\sum_{\sigma\in|\vartheta|}\overrightarrow{\gamma_\sigma} \times \overrightarrow{\delta_\sigma}}{\sqrt{\sum_{\sigma\in|\vartheta|}\overrightarrow{\gamma^2{}_\sigma} \times \sum_{i\in|\vartheta|}\overrightarrow{\delta^2{}_\sigma}}}$$

(4-3)

4.4.1. Sequential Facet

A sequential facet aims to classify data based on the ontological findings of table level clusters, followed by attribute level clusters and then tuple level clusters.

Algorithm 3.1: ***OGDL Sequential Facet Cluster Mapping***

OGDL sequential cluster mapping algorithm

```
Input:
- A set of OGDL clusters $\Sigma \boldsymbol{C_{j,k}} \mid \boldsymbol{j := 1, 2, 3 \ldots n\ and\ k := \{\tau, \partial, \rho\}}$
        $\boldsymbol{\tau:= \text{table level}, \partial:= \text{attribute level}\ and\ \rho: \text{tuple level}}$
- A hash gram matching function: h-gram (see section 3.3 & 3.4)
- A cluster correlation (ICC) function: $\boldsymbol{\Psi}$ (see section 4.3)
- correlation level: $\rho$ // default .75 value set by the user at runtime

Output:
A set of OGDL cluster mappings $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_{n,} \rightarrow \gamma_{n,}})\ \boldsymbol{where\ \rho := stem, \gamma := leaf}$
1.  Initialise empty cluster mapping
    stem($\boldsymbol{\rho_i}$) := {∅}, leaf ($\boldsymbol{\rho_i \rightarrow \gamma_i}$):= {∅}

// sequential facet cluster mapping
2.  FOR ∀ $\boldsymbol{t_i, t_{i+1}} \in \Sigma \boldsymbol{C_{j,k=\tau}} \mid \boldsymbol{\Psi}\ (\boldsymbol{t_i, t_{i+1}}) \sqsubseteq \rho$
3.    FOR ∀ $\boldsymbol{a_{i,t_i}, a_{i+1,t_{i+1}}} \in \Sigma \boldsymbol{C_{j,k=\partial}} \mid \boldsymbol{\Psi}\ (\boldsymbol{a_{i,t_i}, a_{i+1,t_{i+1}}}) \sqsubseteq \rho$
4.      FOR ∀ $\boldsymbol{p_{i,a_{i,t_i}}, p_{i+1,a_{i+1,t_{i+1}}}} \in \Sigma \boldsymbol{C_{j,k=\rho}}$:
5.        IF $\boldsymbol{\Psi}\ (\boldsymbol{p_{i,a_{i,t_i}}, p_{i+1,a_{i+1,t_{i+1}}}}) \sqsubseteq \rho$
6.          $(\boldsymbol{\rho_{i+1} \rightarrow \gamma_{i+1}}) := |\boldsymbol{a_{i,t_i}, a_{i+1,t_{i+1}}}|$
7.        END IF
8.      NEXT
9.    NEXT
10. NEXT

RETURN $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_n \rightarrow \gamma_n})$
```

The multi-faceted clustering process as depicted in Algorithm 3.1 starts with an unorganized cluster set (step 1), and the algorithm commences sequential mapping with empty cluster stem-and-leaf mappings that contain all possible relationships. The sequential faceting (arrangement) process sequentially captures these inter-cluster relationships based on ontological findings of table level clusters (step 2); attribute level clusters (step 3); and lastly of tuple level clusters (step 4). If the clusters are inter-correlated, they are mapped into a stem and leaf representation (step 5-6). An example of sequential facet is depicted in Figure 4.5. As can be seen from the diagram, in a sequential facet, if there are table level cluster matches, we can assume that attributes within these tables could match.

4.4.2. Parallel Facet

A parallel facet does not prioritize any sequence order, and equally classifies data based on the chance of finding pairs within table level clusters; within attribute level clusters or within tuple level clusters.

Algorithm 3.2: ***OGDL Parallel Faceted Cluster Mapping***

OGDL parallel cluster mapping algorithm

Input:

- A set of *OGDL clusters* $\Sigma \boldsymbol{C_{j,k}} \mid \boldsymbol{j} := \mathbf{1,2,3} \ldots \boldsymbol{n}\ \boldsymbol{and}\ \boldsymbol{k} \coloneqq \{\boldsymbol{\tau}, \boldsymbol{\partial}, \boldsymbol{\rho}\}$

  $\boldsymbol{\tau}\mathbf{:=table\ level}, \boldsymbol{\partial}\mathbf{:=attribute\ level}\ \boldsymbol{and}\ \boldsymbol{\rho}\mathbf{:tuple\ level}$
- A hash gram function: *h-gram* (see section 3.3 & 3.4)
- A set of sequential cluster mappings: $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_n} \rightarrow \boldsymbol{\gamma_n})$
- A cluster correlation (ICC) function: $\boldsymbol{\Psi}$ (see section 4.3)
- correlation level: ρ // default .75 value set by the user at runtime

Output:

A set of *OGDL cluster mappings* $\Sigma \boldsymbol{C_n}: \left(\boldsymbol{\rho_{n,}} \rightarrow \boldsymbol{\gamma_{n,}}\right)\ \boldsymbol{where\ \rho \coloneqq stem, \gamma \coloneqq leaf}$

```
// parallel facet cluster mapping
1.  FOR ∀ $\boldsymbol{t_i}, \boldsymbol{t_{i+1}} \in \Sigma \boldsymbol{C_{j,k=\tau}}$
2.    IF $\boldsymbol{\Psi}\ (\boldsymbol{t_i}, \boldsymbol{t_{i+1}}) \sqsubseteq \rho$
3.      $(\boldsymbol{\rho_{i+1}} \rightarrow \boldsymbol{\gamma_{i+1}}) := |\boldsymbol{t_i}, \boldsymbol{t_{i+1}}|$
4.    END IF
5.  NEXT
6.  FOR ∀ $\boldsymbol{a_{i,t_i}}, \boldsymbol{a_{i+1,t_{i+1}}} \in \Sigma \boldsymbol{C_{j,k=\partial}}$ & $\not\mathrm{E}\ (\boldsymbol{\rho_i} \rightarrow \boldsymbol{\gamma_i})$
7.    IF $\boldsymbol{\Psi}\ (\boldsymbol{a_{i,t_i}}, \boldsymbol{a_{i+1,t_{i+1}}}) \sqsubseteq \rho$
8.      $(\boldsymbol{\rho_{i+1}} \rightarrow \boldsymbol{\gamma_{i+1}}) := |\boldsymbol{a_{i,t_i}}, \boldsymbol{a_{i+1,t_{i+1}}}|$
9.    END IF
10. NEXT
11. FOR ∀ $\boldsymbol{p_{i,a_{i,t_i}}}, \boldsymbol{p_{i+1,a_{i+1,t_{i+1}}}} \in \Sigma \boldsymbol{C_{j,k=\rho}}$ & $\not\mathrm{E}\ (\boldsymbol{\rho_i} \rightarrow \boldsymbol{\gamma_i})$
12.   IF $\boldsymbol{\Psi}\ (\boldsymbol{p_{i,a_{i,t_i}}}, \boldsymbol{p_{i+1,a_{i+1,t_{i+1}}}}) \sqsubseteq \rho$
13.     $(\boldsymbol{\rho_{i+1}} \rightarrow \boldsymbol{\gamma_{i+1}}) := |\boldsymbol{p_{i,a_{i,t_i}}}, \boldsymbol{p_{i+1,a_{i+1,t_{i+1}}}}|$
14.   END IF
15. NEXT

RETURN $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_n} \rightarrow \boldsymbol{\gamma_n})$
```

The parallel cluster mapping is depicted in Algorithm 3.2. When the core sequential level mapping has been identified, the algorithm uses parallel faceting to recursively check all potential density reachable cluster mappings. During this process, the algorithm ignores the order of precedence and attempts to find independent matches within table (step 1-5), attribute (step 6-9) and tuple level (step 11-15) clusters. This approach is aimed at mapping different sets of clusters through the density of their ontological relationships, not otherwise identified during the sequential faceting process. Figure 4.5 shows an example of sequential followed by parallel faceting. In a sequential facet, if there are table level cluster matches, we assume that attributes within these tables match. However, in a parallel facet, if table level clusters don't match, the attributes can still match.

4.4.3. Mixed Facet

A mixed facet classifies data through combined cross referencing at the table, attribute and tuple cluster levels.

Algorithm 3.3: ***OGDL Mixed Facet Cluster Mapping***

OGDL mixed facet cluster mapping algorithm

Input:

- A set of *OGDL clusters* $\Sigma \boldsymbol{C_{j,k}} \mid \boldsymbol{j := 1, 2, 3 \ldots n \; and \; k} \coloneqq \{\boldsymbol{\tau, \partial, \rho}\}$
  $\boldsymbol{\tau} := \textbf{table level}, \boldsymbol{\partial} := \textbf{attribute level} \; \boldsymbol{and \; \rho}: \textbf{tuple level}$
- A hash gram function: *h-gram* (see section 3.2)
- A set of sequential and parallel cluster mappings: $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_n} \rightarrow \boldsymbol{\gamma_n})$
- A cluster correlation (ICC) function: $\boldsymbol{\Psi}$ $(\textbf{see section 4.3})$
- correlation level: ρ // default .75 value set by the user at runtime

Output:

A set of *OGDL cluster mappings* $\Sigma \boldsymbol{C_n}: \left(\boldsymbol{\rho_{n,}} \rightarrow \boldsymbol{\gamma_{n,}}\right) \boldsymbol{where \; \rho \coloneqq stem, \gamma \coloneqq leaf}$

```
// mixed facet cluster mapping
1. FOR $\forall \boldsymbol{m_i, n_{i+1}} \in \Sigma \boldsymbol{C_{j,k=\{\tau,\partial,\rho\}}}$ & $\nexists (\boldsymbol{\rho_i} \rightarrow \boldsymbol{\gamma_i})$
2.    IF $\boldsymbol{\Psi} (\boldsymbol{m_i, n_{i+1}}) \sqsubseteq \rho$
3.       $(\boldsymbol{\rho_{i+1}} \rightarrow \boldsymbol{\gamma_{i+1}}) := |\boldsymbol{m_i, n_{i+1}}|$
4.    END IF
5. NEXT

RETURN $\Sigma \boldsymbol{C_n}: (\boldsymbol{\rho_n} \rightarrow \boldsymbol{\gamma_n})$
```

Time-series data is common and pervasive in many applications – it is thus a very important issue to deal with, which merits special attention [18]. Algorithm 3.3 shows the mixed facet cluster mapping technique. In contrast to sequential and parallel faceting, mixed faceting aims to discover inter-cluster relationships by cross-referencing clusters in different dimensions. The algorithm begins by identifying and attempting to find matches across table, attribute and tuple level clusters (step 1). If the matches are found, which are not identified in the previous stages, the matching clusters are recorded (step 2-5). The intuition for this is based on heuristics and time-series based cluster mapping. In a mixed facet, a sequence of time-series values can be represented by associated attributes. For instance, in the World Bank [24] database, the tuple values from the Health Nutrition and Population Statistics table included {1961M01, 1962M2, 1963M3,…}, which had to be matched with the Global Economic Monitor Terms of Trade table attribute names {1960, 1961, 1962,…}.

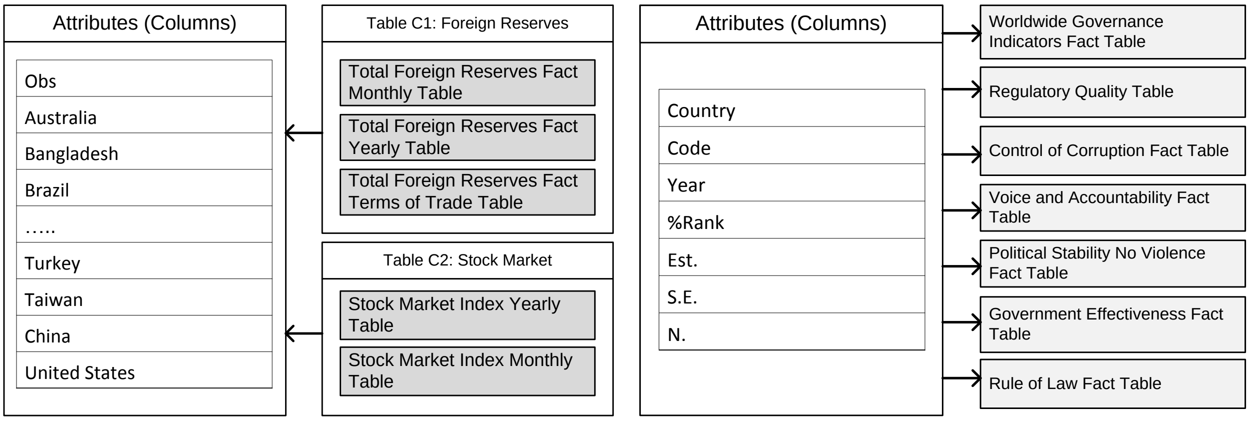

Attributes (Columns)
Obs
Australia
Bangladesh
Brazil
…..
Turkey
Taiwan
China
United States
Table C1: Foreign Reserves
Total Foreign Reserves Fact Monthly Table
Total Foreign Reserves Fact Yearly Table
Total Foreign Reserves Fact Terms of Trade Table
Table C2: Stock Market
Stock Market Index Yearly Table
Stock Market Index Monthly Table
Attributes (Columns)
Country
Code
Year
%Rank
Est.
S.E.
N.
Worldwide Governance Indicators Fact Table
Regulatory Quality Table
Control of Corruption Fact Table
Voice and Accountability Fact Table
Political Stability No Violence Fact Table
Government Effectiveness Fact Table
Rule of Law Fact Table

Figure 4.5: An example of (a) sequential and (b) parallel facet matches from The World Bank [24]

4.5. Transforming Cluster Correlations into Schema Mapping

Creating schema structures require the identification of candidate; primary; and foreign key relationships. In the previous section, we performed inter-clusters mapping, using the OGDL framework to perform incremental pair-wise comparison between attributes in each table: Semantic cluster mappings determine possible candidate keys but this does not give us information regarding their relationships and directions.

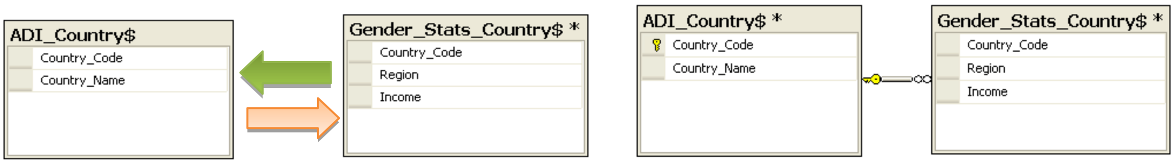


Figure 4.6: An example of finding primary/foreign key relationships between different attributes based on cluster density relationships

In order to achieve this, we take our approach a step further and expand our framework by consequently transforming the previous cluster correlation results into global schema structures, with a unified representation of the entire dataset. The rationale behind our approach is that when two clusters are strongly correlated, we can use these clusters to predict primary/foreign key relationships. We do this by computing the density of relationships between clusters with one attribute (a.k.a. independent clusters) and clusters with a second attribute (a.k.a. dependent clusters), as shown in Figure 4.6.

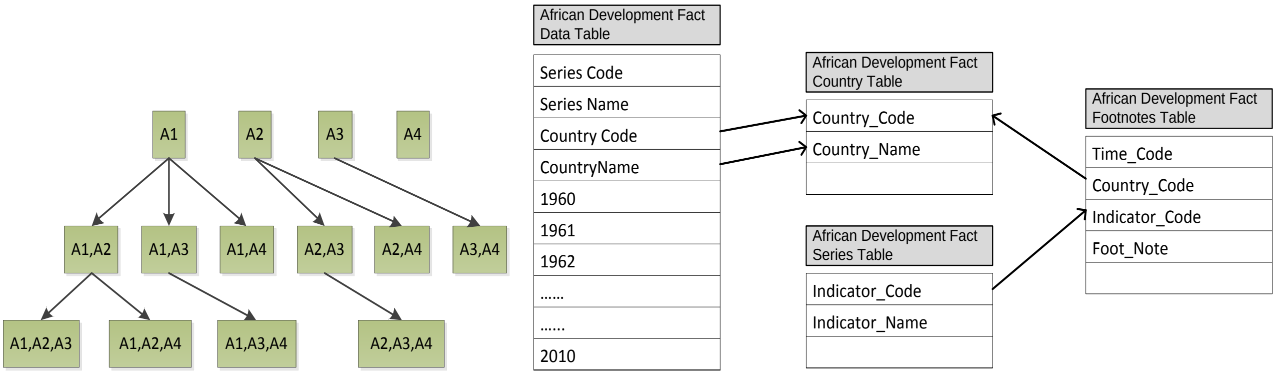


Figure 4.7: (a) Incremental pair-wise comparisons applied on cluster pairs in order to determine primary keys in a table. (b) an example of parent-child relationships extracted from the African development indicators dataset (The World Bank [24])

If Ω is the parameter specified by the user and we have $A_{1,}A_2, A_{3,}\ and\ A_4$ attributes (as shown in Figure 4.7(a)), then incremental pair-wise tuple mapping sons $\{A_1, A_2\}, \{A_1, A_3\}, \dots \{A_3, A_4\}, \dots \{A_2, A_3\ A_4\}$, to the count of the Ω-item set, are performed to discover candidate keys. Consequently, potential primary/foreign key relationships are identified by computing the density of the relationships between cluster pairs of different table attributes. The intuition is that high densities of relationships identify potential primary/foreign key relationships. This computation uses the multiplicity property, which for our purpose is defined as the maximum number of child table cluster entities $\cup A$ able to link with the parent table cluster entities $\cup B$, in order to form parent-child relationships, as depicted in the below equation.

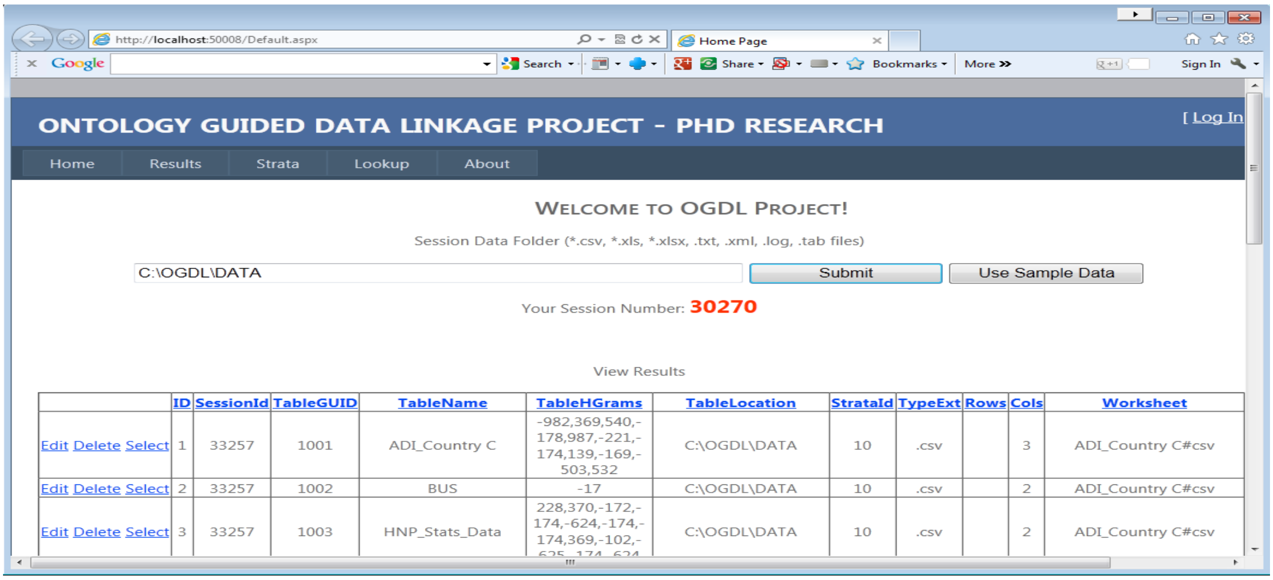
$$if\ A \subseteq B, then\ \cup A \subseteq \cup B\ , otherwise\ \cup B \subseteq \cup A$$

(4-4)

Figure 4.7(b) shows an example of parent-child (primary-foreign) relationships, as data linkages between African development indicator tables from The World Bank database [24]). If no primary-foreign key relationships between clusters are found, the matching clusters tend to partially match, and we consider them as partially related keys because of their partial level of significance. Despite these relatively weak relationships, our experiments have shown that these partial relationships, currently ignored by existing techniques, can contain significant data facts.

4.6. Experimental Evaluation

Many organizations freely share their real-world data (as data files, datasets, data models, etc.) on the internet. These data are mostly in third-normal form, but the lack of a 'gold standard' for data linkage represents one of the major challenges in evaluating such real-world data collected from multiple domains. Towards this purpose, we quantify the benefits of our proposed framework and measure the sensitivity of our framework results using a 10-fold cross validation approach. It is important to note that our algorithms are not a panacea for the entire data linkage process, and that we aim to limit the amount of manual feasibility checks necessary to cross-validate the results. We presented a new approach to the discovery of semantic attribute correlations by multi-faceted identification and analyses of ontologies used in unlinked real-world data. While sophisticated enterprise data mining tools already exist, we have presented a much simplified approach to data linkage for the purpose of finding accurate crucial data facts. Our approach provides better results than comparable approaches and adds value by quantifying the expected computer throughput.



ONTOLOGY GUIDED DATA LINKAGE PROJECT - PHD RESEARCH

WELCOME TO OGDL PROJECT!

Session Data Folder (*.csv, *.xls, *.xlsx, .txt, .xml, .log, .tab files)

C:\OGDL\DATA

Your Session Number: 30270

View Results

	ID	SessionId	TableGUID	TableName	TableHGrams	TableLocation	StrataId	TypeExt	Rows	Cols	Worksheet
Edit Delete Select	1	33257	1001	ADI_Country C	-982,369,540,-178,987,-221,-174,139,-169,-503,532	C:\OGDL\DATA	10	.csv		3	ADI_Country C#csv
Edit Delete Select	2	33257	1002	BUS	-17	C:\OGDL\DATA	10	.csv		2	ADI_Country C#csv
Edit Delete Select	3	33257	1003	HNP_Stats_Data	228,370,-172,-174,-624,-174,-174,369,-102,-625,-174,-624	C:\OGDL\DATA	10	.csv		2	ADI_Country C#csv

Figure 4.8: OGDL Prototype Interface

4.7. Prototype Development

We monitored our experimental results by validating the accuracy, performance and scalability of our proposed framework as benchmarks for the evaluation process. We conducted our experiments by developing and using three prototypes. First, we developed the core 'OGDL Data Miner' project as shown in Figure 4.8 which performs the bulk of our proposed framework tasks at different stages. We then developed the 'OGDL Cluster Search Engine', an interactive and user-friendly tool to visualize the cluster stem-and-leaves formed across multi-domain databases. By clicking on any searchable cluster, the user can drill into its correlated clusters for knowledge discovery and for exploration of 'chains-of-facts'. We also developed the 'OGDL Performance Monitor' tool to analyze the scalability of our framework while running on different machines.

All our prototypes have been developed using Microsoft Visual Studio 2008 in C# & .Net with Microsoft SQL 2008 as the backend database. A demo version of our prototype is available for researchers and other staff to conduct experiments and add their valuable feedback, as we continue to improve the performance of our system. All experimental results are exportable in XRFF (eXtensible attribute-Relation File Format), which is an XML-based extension of the ARFF format [17]. Since performance measurements are given in absolute times, it is necessary to take the computing environment into account. Our experiments have been conducted on a Windows machine (CPU @ 2.0GHz, RAM 8 GB, and 64-bit OS).

In order to avoid any deceptively fast performances due to caching, we flushed the memory cache (system and query cache) prior to conducting each experiment. All our data linkage operations were performed in local memory and the final results were written into the backend DBMS. These include the findings of clusters and cluster pairs, and the calculations involved in finding attribute correlations and schema structures. Multiple time stamps have been added in order to record the performance of our experiments at different levels. Our experimental results have been graphically analyzed using ASP.Net 3.0 Charts and the Microsoft Reporting service (in MS Visual Web Developer 2010).

Table 4.1: Experimental dataset characteristics

Dataset#	Tables#	Attributes#	Tuples#
The World Bank Data Catalog [24]	4512	67680	2.0 M
The US Federal Govt. Data Catalog [30]	3155	43339	1.4 M
National Climatic Data Center [49]	255	1350	15 K

4.8. Data Setup

We evaluated our OGDL framework using datasets collected from a variety of real-world data sources, with the intuition that the lack of a common domain and the inconsistencies in the data would effectively model the current research gap of multi-domain experimentation. The characteristics of our experimental data are shown in Table 4.1. The World Bank [24] datasets includes varied statistical information about country level development, as part of the World Bank's mission to alleviate poverty. The US Federal Government Datasets [30] provides a broad range of government related data to evaluate the efficiency and effectiveness of various governments. The datasets also includes US relationships with other countries in regards to various global issues. The National Climatic datasets [49] holds the world's largest data archives on climate information.

To validate the cost analysis of our OGDL framework and to measure the accuracy of our approach, we conducted a series of experiments with different stratified sample sets of data (see section 4.2) and studied the result averages. Our best-case scenario was to discover relationships across multiple datasets, whereas our worst-case scenario was to find attribute relationships within identical databases (i.e. every column matches with one other column).

4.9. Evaluation Principles

The lack of a gold standard format for integration of real-world datasets poses a significant challenge to accurately evaluate data linkage performance. Hence, we focused on obtaining accurate results prior to performing validation against OGDL measurements. We developed a fully supervised Brute Force (BF) attribute based matching technique to evaluate the accuracy of attribute pairs across unrelated datasets. We used the BF technique to extensively calculate exact domain overlaps that considers all tuples, through determination of the weights of join attributes. The BF method determines the correlation between every pair of attribute relationships. The results obtained from the BF approach were used as a new ‘Gold Standard’ to evaluate our earlier OGDL measurements. The earlier measurements originated from analyses of different horizontal and vertical data subsets, as determined by multi-faceted multi-dimension relationships (see section 4.3 and 4.6).

We conducted two different types of experiments on each database, which we coined as ‘WITHIN’ and ‘BETWEEN’ comparisons. ‘WITHIN’ experiments aim to test the worst-case scenario (as described before) by calculating OGDL accuracy and performance when applied to discover data linkages between identical databases. We assume that every column matches at least one other column as its counterpart in identical databases. ‘BETWEEN’ experiments aim to test the best-case scenario (as described before) by calculating OGDL accuracy and performance when applied to discover data linkages between variable databases. Each experiment was repeated 10 times and the statistical mean values were recorded in our experimental graphs. We measured the accuracy, performance and scalability of our results in comparison to the new 'Gold Standard' (GS).

4.9.1. Accuracy Metrics

Accuracy tests were conducted to ensure that the obtained results were closely associated with the expected true values. We first determined the formation of different sets of clusters using identical databases and then compared them to the 'Gold Standard'. We measured the cluster formation accuracy A as depicted in the below equation.

$$A = \frac{\#\ of\ correct\ clusters\ detected}{\#\ of\ GS\ clusters\ evaluated}\ X\ 100\% \tag{4-5}$$

Figure 8(a) shows the associated OGDL cluster formation percentages in different databases and Figure 4.9(f) shows the OGDL cluster counts in comparison to that of the GS. We note that the number of clusters discovered by OGDL is very close to the expected true values. Figure 4.9(b) shows the number of errors that occurred in different data sample dimensions. We notice that the OGDL approach can identify significant attribute relationships with minimal errors even with small data sample sizes,. Further evaluation of the effectiveness of OGDL algorithms when applied to different databases included calculations of different levels of correlation strengths. We used false positive (type 1 error) and false negative (type 2 error) correlations, which are usually applied to test statistical hypotheses. A false positive correlation is defined as an erroneously defined correlated relationship; and a false negative correlation is defined as a correlated relationship that is erroneously not defined. The results are shown in Figure 4.9(e). The error percentage ‘E’ is calculated by taking into account both the OGDL attribute correlation errors W_{OGDL} and the GS attribute correlation errors W_{GS}, as shown in the below equation.

$$E = \frac{|W_{GS}(A_1,A_2) - W_{OGDL}(A_1,A_2)|}{W_{GS}(A_1,A_2)} \tag{4-6}$$

Table 4.2: Accuracy measurement table showing precision and recall values for top-10 table level, attribute level and tuple level clusters

Cluster	Cluster Size	Precision (p)	Recall (r)	F Score
***Top-k* Table Level Clusters**				
Agriculture	102	91	92	91.49
Economic Policy	20	73	84	78.11
Education	569	95	99	96.95
Energy	42	61	92	73.35
Health and Nutrition	158	93	84	88.27
Financials	458	85	86	85.49
Labor Force	62	91	82	86.26
Poverty	85	93	84	88.27
Foreign Aid	77	95	76	84.44
Science and Technology	258	79	78	78.49
		85.6	85.7	85.11
***Top-k* Attribute Level Clusters**				
Amount	102	95	87	90.82
Geography	220	93	78	84.84
Year	543	84	77	80.34
Name	52	57	86	68.55
Indicator	856	88	67	76.07
State	1027	85	85	85.00
Quantity	543	81	98	88.69
Value	345	92	97	94.43
Certificate	54	98	67	79.58
Note	34	94	74	82.80
		86.7	81.6	84 84.11
***Top-k* Tuple Level Clusters**				
2500	455	87	76	81.12
Aggregates	54	99	56	71.53
Adult Literacy	64	89	56	68.74
Nelson	65	86	88	86.98
Airplane	764	69	78	73.22
United States	867	95	56	70.46
Migration and Refugee	97	78	86	81.80
Child Survival	945	98	67	79.58
Accident	543	67	88	76.07
University	434	59	93	72.19
		82.7	74.4	76.17

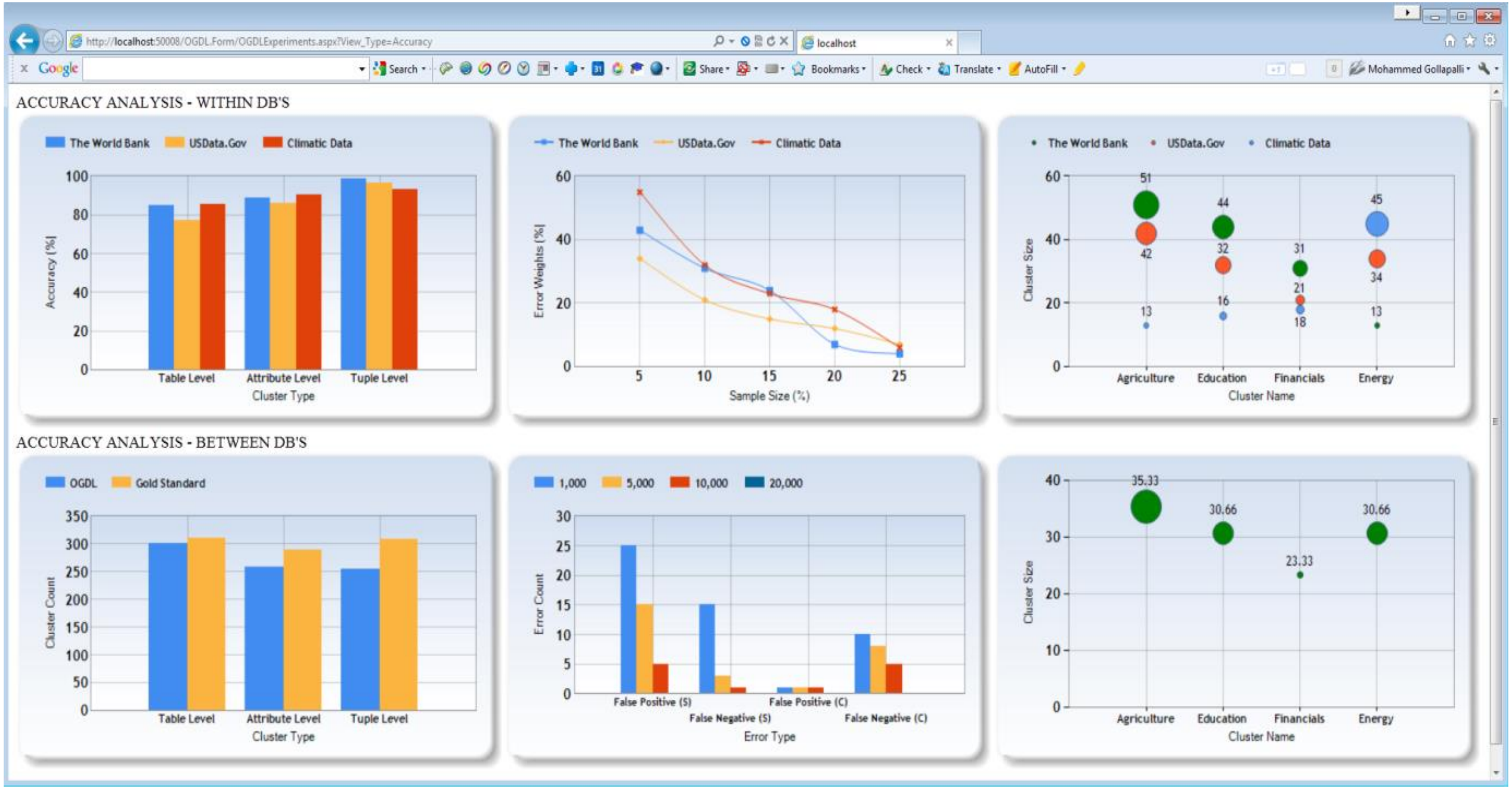


Figure 4.9: Clockwise starting from top left (a) accuracy comparison with the gold standard in individual databases; (b) effect of strata size percentage against error weight percentage in individual databases; (c) formation of different clusters and their sizes in individual databases; (d) formation of different clusters and their sizes on combined databases; (e) false positive and false negative correlation results on different strata sizes and (f) accuracy comparison to the gold standard on combined datasets.

We examined, as the strata size varied, the changes to the ratio of correlation of true matches. The results verified that the OGDL approach yields an acceptable percentage of expected results, even with small strata samples. We evaluated the 'OGDL Clustering' algorithm by measuring the quality of the different sets of clusters that were formed. The quality of our results was monitored through experimentation on different databases and by recording 'precision' and 'recall' measurements. Precision *p* and recall *r* is calculated as shown in the below equation.

$$p = \frac{|\{GS\ based\ relevant\ clusters\} \cap \{OGDL\ Clusters\ retrieved\}|}{|\{OGDL\ clusters\ retrieved\}|} \quad r = \frac{|\{GS\ relevant\ clusters\} \cap \{OGDL\ clusters\ retrieved\}|}{|\{GS\ relevant\ clusters\}|}$$

(4-7)

Table 4.2 includes the percentages of the top 10 table level precision (p) and recall (r) values, where attribute and tuple level clusters were formed using the OGDL technique. The precision column shows the percentage of different sets of clusters that was correctly determined by the OGDL framework as belonging to the same group, compared to the gold standard. The recall column shows the number of clusters that were accurately identified, as a percentage of the total number of cluster elements. We observe that the employment of the OGDL clustering technique is associated with a higher precision and less recall than existing techniques, due to its sequential and iterative application to increasingly similar clusters. A detailed study of our results demonstrated that the OGDL findings include complex attribute relationships even for attributes with different patterns.

We observe that clusters with low precision, such as 'Economic Policy' and 'Energy' occur infrequently due to the scarcity of data correlated to these clusters, and these clusters have little significance. Figure 4.9(c) and Figure 4.9(d) shows the accurate number of cluster overlaps formed on different databases, as a clear indication of the effectiveness of the OGDL clustering technique. Table 4.2 also shows the F measure results, which is the harmonic mean that considers both precision and recall scores. A 'F measure' with a score of 1 represents the best case scenario, while 0 represents the worst case scenario. The F measure, using precision (p) and recall (r), is depicted in the below equation.

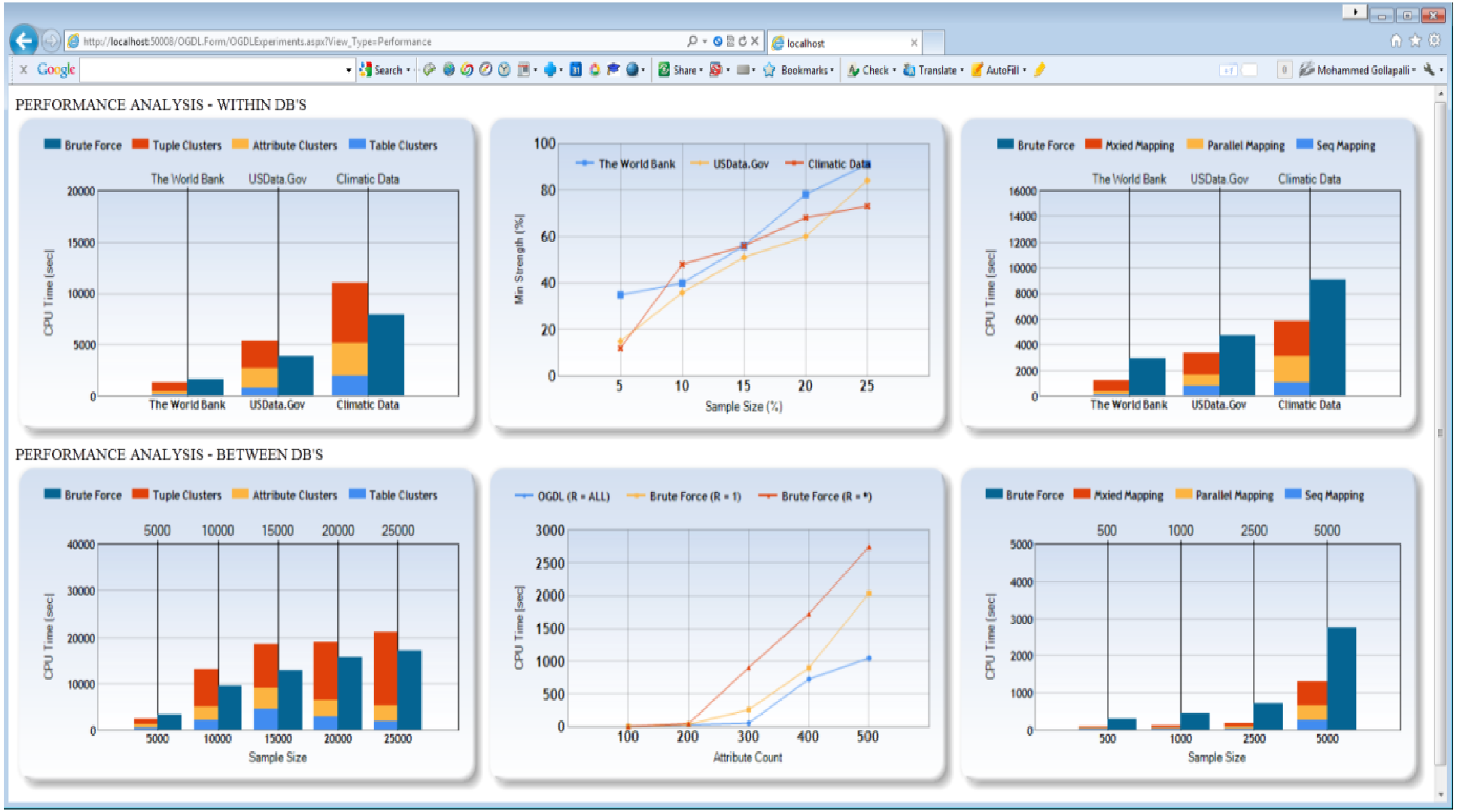
$$F\ score = 2.\frac{p.r}{p+r}$$

(4-8)

4.9.2. Performance Metrics

The analysis of the performance of the OGDL framework has included the extensive evaluation of its algorithms, applied to different horizontal and vertical data subsets. We compared the CPU time for OGDL table, attribute and tuple level clustering with that of using the BF approach. Figure 4.10 (a) shows the runtime cost associated with cluster formation in different databases. We observe that the OGDL clustering approach scales gracefully when compared with the BF clustering approach. Although OGDL clustering at the tuple level seems to take longer, thus downgrading its overall performance in comparison to the attribute-based BF clustering approach, this is outweighed by the significant performance gains reaped during the actual cluster mapping process, as shown in Figure 4.10 (c) for individual databases and in between databases as depicted in Figure 4.10 (d). The performance gains of our framework ranged between 20% and 38% on different databases. The reason that our framework performs so well is due to its systematic multi-faceted mapping stages (see section 4.4) that maps cluster concept trees. Similar experiments to perform attribute matching have been conducted between different datasets, and as can be seen in Figure 4.10 (d), the CPU time cost is significantly less using the OGDL approach than when using the BF approach.

We further evaluated the performance of the OGDL framework by monitoring the sample size requirements for discovering cluster mappings in different databases. As can be seen in Figure 4.10 (b), OGDL is able to discover a minimum of 60% of relationships correlations between attributes, using as little as 10-15% of the total dataset. The attribute correlations that were found were the same as those found by performing multi-faceted cluster mapping. This further proves that OGDL is a fast learning tool that can significantly contribute to address this research problem.



PERFORMANCE ANALYSIS - WITHIN DB'S

PERFORMANCE ANALYSIS - BETWEEN DB'S

Figure 4.10: OGDL Performance Analysis. Clockwise starting from top left (a) performance comparison with BF in different databases; (b) effect of data sample size on correlation strengths within different databases; (c) CPU run time costs for OGDL vs. BF relationship mapping within individual databases; (d) CPU run time cost associated with OGDL vs. BF cluster mapping of multi-domain data; (e) CPU time when running OGDL and BF algorithms for finding relationships, as a function of the number of attributes; (f) CPU time comparison for the OGDL Clustering vs. the BF Clustering technique.

We monitored the OGDL technique's performance in terms of discovering relationships when applied to different attribute sizes. We used incremental attribute counts (100, 200, 300, etc.) as shown in Figure 4.10 (e). The experimental results demonstrated that OGDL scales almost linearly with an increasing attribute count. We observe that our approach identifies the expected number of attribute relationships by an order of magnitude, and that its performance is much better than that of the BF approach, when applied for one-to-one or one-to-all relationship finding. Figure 4.10 (f) shows the execution times of OGDL for table; attribute; and tuple level clustering when applied between three disparate experimental databases each with an increasing number of sample size. We observe that table and attribute level clustering is discovered within a reasonable amount of time. Furthermore, we observed that tuple level clustering also scales gracefully with increasing data sample size.

The application of our OGDL approach is associated with significant performance measures, using as little as 10% of the sample size. In other words, related performance gains are dependent on the size of the strata sample and the number of attributes. OGDL facet mapping determines the majority of core relationships, thus increasing the effectiveness of our system as shown in Figure 4.10 (d). After having conducted this experiment repeatedly, we conclude that the multi-faceted mapping approach is highly effective for discovery of the fine structures of individual and group datasets.

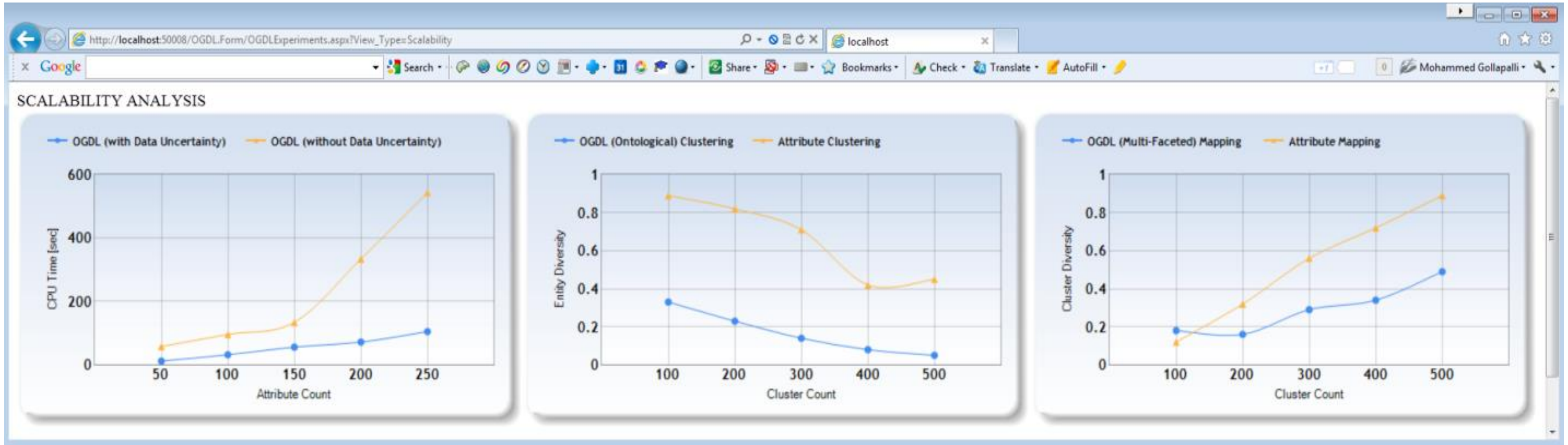
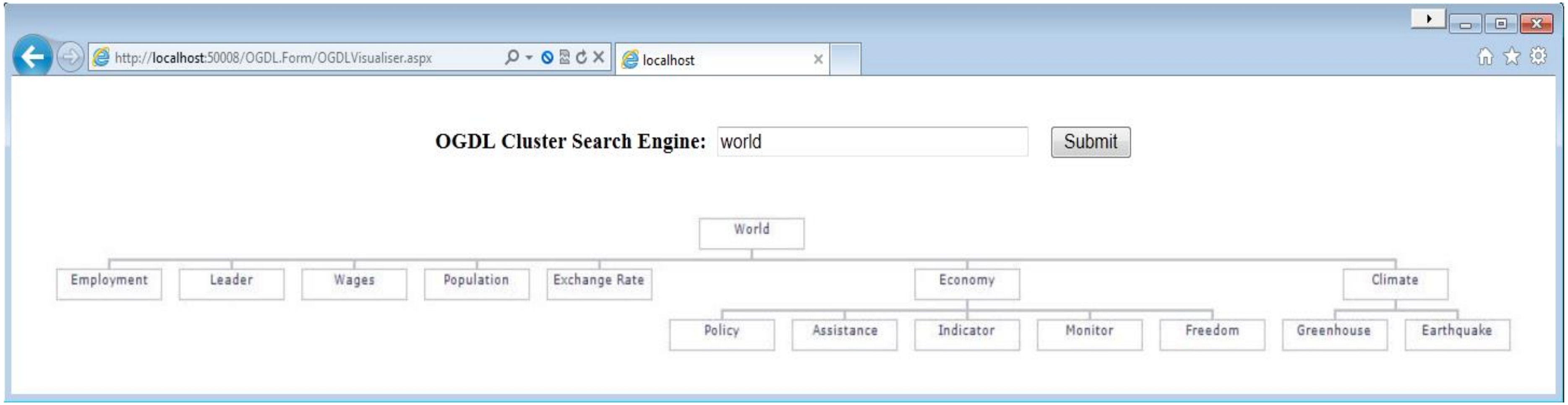


Figure 4.11: (a) OGDL clustering with and without the data uncertainty process; (b) entity dispersion results for OGDL vs. BF (attribute based) clustering; (c) cluster diversity results for OGDL vs. BF (attribute based) mapping



OGDL Cluster Search Engine: world

Figure 4.12: The OGDL cluster search engine browses and visualizes clusters formed with stem-and-leaf relationships

4.9.3. Scalability Metrics

Scalability tests using the OGDL approach have been conducted to assess the workload required; related system result throughputs; and the ability to handle varied sets of data. The scalability tests in our experiments are focused on quantifying the ‘entity dispersion’ and ‘cluster diversity’ measurements at different levels. An entity dispersion measurement is applied on the OGDL clustering technique to determine the frequency of occurrence of entity best fitting, as the number of clusters increase. Lower dispersion means that the entity best corresponds to a cluster with similar cluster items (based on ontological similarities), and higher dispersion means that the entity can erroneously be assigned to unrelated clusters.

Figure 4.11 (a) shows the importance of applying the ‘data uncertainty’ process (see section 4) prior to running the OGDL technique. As can be seen from the results, running OGDL algorithms can get extremely expensive if the input data is not properly classified. Figure 4.11 (b) shows the entity dispersion graph for OGDL clustering and for BF (attribute based) clustering. The best entity dispersion (ideally in primary key attribute fields) has a dispersion value of 0, and the best cluster diversity has a diversity value of 1. We observe that OGDL clustering performs better with lower dispersion, and outperforms attribute based clustering. This is due to the significant gains achieved by prior ontology matching at multiple levels. Cluster diversity measurements quantify different cluster mapping relationships, which have previously correctly been identified as expected attribute pairs. Figure 4.11 (c) shows cluster diversity measurements collected through OGDL and BF (attribute based) mapping techniques. As can be seen, OGDL outperforms BF based attribute mapping. In contrast to BF based attribute mapped clusters, OGDL cluster sizes are significantly larger and each cluster represents greater sets of identical and similar entities. This further helps to speed up the mapping process, especially when matching different datasets.

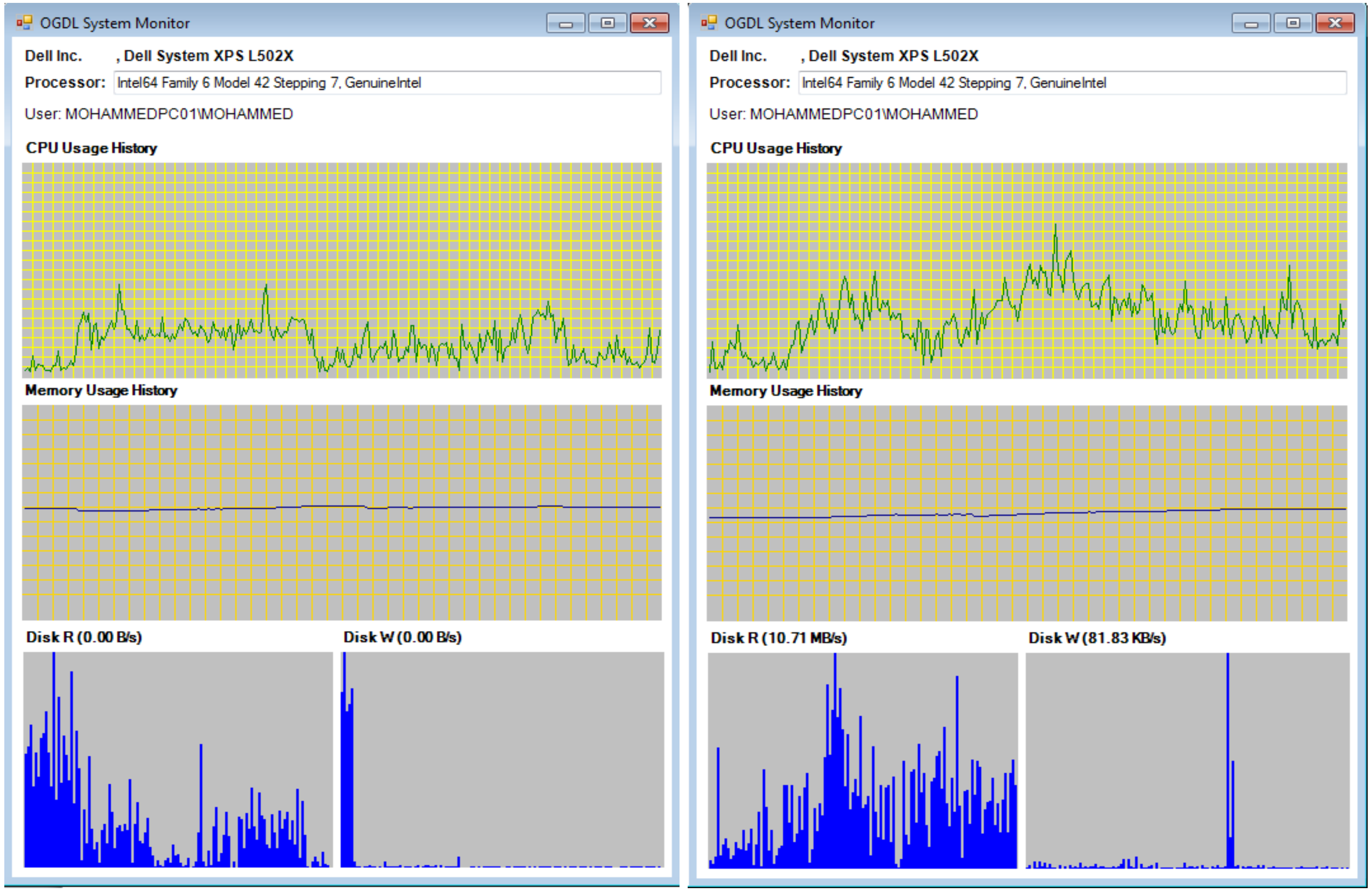
4.10. OGDL Search Engine

Our experimental results have demonstrated that OGDL clustering is able to discover ontology based duplicates, which is an existing research problem when using fixed threshold attribute based clustering approaches. We believe that our framework can significantly improve fact finding and knowledge discovery measurements by the employment of the h-gram matching [79] technique in our OGDL algorithms. Figure 4.12 shows the 'OGDL Cluster Search Engine' prototype that was developed to visualize and validate the OGDL cluster mapping algorithm. We studied the effect of choosing clusters at different levels of the OGDL cluster tree. Having searched for a data fact, OGDL search results are displayed down to the lowest level of the cluster and include relationships with other leaf clusters. As can be seen from Figure 4.9 (a), the accuracy percentages of the OGDL framework are very close to the expected true values as per the GS. The accuracies of our framework for different databases ranged from 77% to 98%. Our results vary slightly in comparison to the GS results due to the stratified sampling process employed during the 'Data Uncertainty' stage (refer to section 4). Similar experiments were conducted to determine the percentages for different sets of clusters that were formed between unrelated databases.

After having conducted rigorous testing, we observed that OGDL clustering is significantly faster and highly accurate in comparison to existing techniques. OGDL clustering results are at least 10% more accurate than BF (attribute only based) clustering results. This further demonstrates that OGDL mapping has the capacity for learning the finer structures between different databases. BF is considerably slower and less accurate than OGDL. BF leaves many attributes unmapped and unassigned to semantic equivalent clusters. We found BF to be accurate when applied to The World Bank dataset [24], which has normalized formatted data and is thus easily simplified for the clustering process. However, BF clustering is not a suitable technique to map complex data clusters such as that of the National Climatic dataset [49]. We believe that the OGDL technique is an effective tool for accurate data linkage mapping of complex data clusters and that it can easily be applied to open source and commercial datasets from many different domains. We also believe our approach is considerably better than those used by enterprise mining systems such as SAP, which are costly, complicated and necessitates specialized IT knowledge and skills. Our experiments also established the usefulness of the OGDL technique for modal and feature selection.

To further demonstrate and compare hardware and software costs, we monitored the CPU, memory and hardware costs associated with OGDL and BF applications during experiments conducted using the windows server[2]. The results obtained through this experimentation were tracked, using our OGDL Performance Monitor prototype tool. Figure 4.13 shows the scalability results as diagnosed on the runtime machine when running OGDL and BF algorithms separately. While both of them do significantly better in terms of memory usage than other similar techniques, OGDL uses the least CPU; i.e. the OGDL framework performs data linkages with less CPU expenditure and with moderate memory usage when compared to the BF approach. This further illustrates that the BF approach of attribute level mapping is very expensive both in terms of hardware and runtime costs. We also believe that OGDL is likely to perform well in other operating systems, with moderate associated hardware requirements. As these results are significant, we conclude that our framework scales well in terms of hardware and software costs.

[2] Running OGDL methods simultaneously on multi-core processors in parallel can further improve the performance.



OGDL System Monitor

Dell Inc. , Dell System XPS L502X

Processor: Intel64 Family 6 Model 42 Stepping 7, GenuineIntel

User: MOHAMMEDPC01\MOHAMMED

OGDL System Monitor

Dell Inc. , Dell System XPS L502X

Processor: Intel64 Family 6 Model 42 Stepping 7, GenuineIntel

User: MOHAMMEDPC01\MOHAMMED

Figure 4.13: OGDL System Monitor tool displaying the CPU, Memory and Hardware disk (read/write) costs in running OGDL (left diagram) and BF (right diagram) methods

4.11. Summary

In this chapter, we first introduced the ‘data uncertainty’ concept that necessitates robust cleaning and automatic data categorization prior to running the bulk of the data classification processes. By performing this step firstly in the OGDL approach, we provided the means for tightly integrating attributes based on ontological domain information; and we introduced a simple unified learning model that can tag frequently occurring clusters. We then presented a practical method for discovering multi-layer ontological data clusters to support practical and crucial information extraction tasks. Through its evaluation, we observed that our OGDL algorithms are fast learners and that they gain maximum accuracy with a small sample of strata sets. We continued by proposing a multi-faceted mapping algorithm for learning the structures of input data from multiple knowledge domains. This method enables the cluster mapping of hierarchical tree structures as concept maps. Given our encouraging evaluation results, we believe that our approach performs accurate attribute level matching, unlike CORDS [16] approach, which rely on heuristics to control the order in which the attributes are mapped; and which exploits domain information by using only a subset of the information. We have also included an explanation of how our results can easily be integrated with IBM or Microsoft’s QBE (Query-by-Example) tools, in order to perform semantic queries. This is part of our future work in consideration.

Our purpose in incorporating indexing techniques was to reduce runtimes while maintaining a high cluster quality. Through extensive evaluation on real-world data, we have shown that the OGDL framework approach can discover data linkages when databases do not share common attributes. Extensive comparative BF testing has been conducted in order to evaluate the accuracy and performance of our framework against the attribute pair matching approach. Our experimental results have shown that OGDL yields accuracies ranging from 84% to 86%, between databases, and up to 93% within individual databases. Through multi-domain experimentation, we have proved that OGDL can be used during the crucial data mining phase for automated graphical analysis and cluster visualization. Our results have also demonstrated that the OGDL algorithms can perform accurate data linkages with as little as 10% of the actual database size that is available, for training purposes. We aimed at enabling our framework to be applied by existing QBE based tools such as IBM or Microsoft’s Data Analyser, to perform sensible queries to support accurate fact extraction, and to support a wide-variety of data-warehouse tasks. We believe that the OGDL approach is an effective approach for practical information extraction and crucial fact finding purposes, and that it performs better than other attribute-based clustering approaches with the same aim. In the future, we plan to work on multiple clustering consolidation and cluster scaling techniques. Future work in this area includes the conduction of runtime tests for fast cluster browsing and to meet semantic reasoning learning needs, compatible with Google style browsing.

In the next chapter, we will introduce extension of OGDL Framework for continued collaborative development and application of our 'Gold Standard' as a semantic reasoning Upper Ontology in a problem-solution learning framework. This will support data integration in a knowledge repository with greatly enhanced data mining capacity, and will enable user-friendly First Order Logic querying to extract meaningful facts without expert IT knowledge and skills. We applied our prototype application to the field of risk management, historically hampered by disparate domain ontologies and datasets. This extension of our OGDL Framework incorporates the capacity to:

- Help develop the Upper Ontology and Learning Framework linkages of their de-identified datasets;
- Perform shared view risk modeling with drilling capacity of time complex, time-stamped datasets that have automated access to machine learning of knowledge repository patterns and anomalies;
- Use interactive decision-tree visualization weighted by ranked levels of evidence and expert agreement to support and accelerate progress towards approximated consensus on findings;
- Store evidentiary time-stamped data mining views referenced as chains-of-relationships towards consensus development;
- Link and rank resources with chains-of-relationships for digital streaming purposes to support quality risk management;
- Compare existing risk indicators and related resource implementation standards and task guidelines based on machine learning results and expert best practice evidence weighting using a Likert scale;
- Label and integrate streaming real-world data in the knowledge repository;
- Choose a variety of data views, including graphical computed windows of opportunity for risk management as per set of risk factors in time and per available resources; GIS linked streamed tracking of risk interventions; virtual team meeting

spaces; and dashboard views with automated pre-defined as well as machine learnt alerts;

- Recommend improved standards and task guidelines per risk profile, and address risk prevention and mitigation for geographic clusters of risk profile attributes, including workforce planning and human resource strategies;
- Semantically reason using the Upper Ontology in the Learning Framework as a regular expression language with First Order Logic capacity to help drive research and resource implementation effort for optimal and sustainable impact.

In conclusion, our framework introduced in this chapter can be used with unnormalized, semi-normalized and normalized databases of various sizes. Through the evaluation of the accuracy, performance and scalability of our framework when applied to unrelated databases with different horizontal and vertical subsets, we proved that the OGDL approach achieves high quality results and that the development of our framework is highly significant, and an important step towards user-friendly semantic reasoning functionality in a Semantic Web environment.

Chapter 5
Extending OGDL Framework for Clinical Success Indicator Development

5.1. Problem Description

Clinical risk management is a complex problem, with stakeholders that include health improvement and service funding sources, patients, carers, and service providers, each with their own silo of disparate health data: 'There is a need to pool together collective knowledge and experience, and infuse it into a decision-support system (DSS) on an ongoing basis' [83, 84]. Evidence of this problem was established by a review of the quality of clinical guidelines in hospitals in Australia, Indonesia, Malaysia, the Philippines and Thailand spanning the period 1988-98 [86]. Australian studies show that around 50% of medication errors occur at interfaces of care [87, 89]; 2–5% drug charts typically contain prescribing errors, and up to 70% of medicines administered intravenously have one or more clinical errors [6]; and that up to 30% of Australian hospital admissions of patients older than 75 years are medication related. [87]. Clinical guideline development has also been found to be enormously time, skill and resource intensive, and there is a general consensus towards improved development of evidence-based clinical practice guidelines in [92].

There has been consistent reference made in Australia to the need for e-health as an approach to improve clinical outcomes, by providing decision-support at the point of care [89, 91, and 93]. It is part of the national Australian e-health strategy to ‘Improve the quality, safety and efficiency of clinical practices by giving care providers better access to consumer health information, clinical evidence and clinical decision support tools’ [89]. Clinical risk management research currently uses the approach of scientific clinical trials, and as a result evidence for improving patient outcomes is limited to the number of patient cases included in such trials [86, 94, and 96]. To make things worse, such data is usually analysed in terms of failure, not success. Also, we have found that published clinical guidelines are not transparently evidence-based or holistic, and that they cannot be tailored for complex patient problems at the care interface [86, 87, 92, and 93].

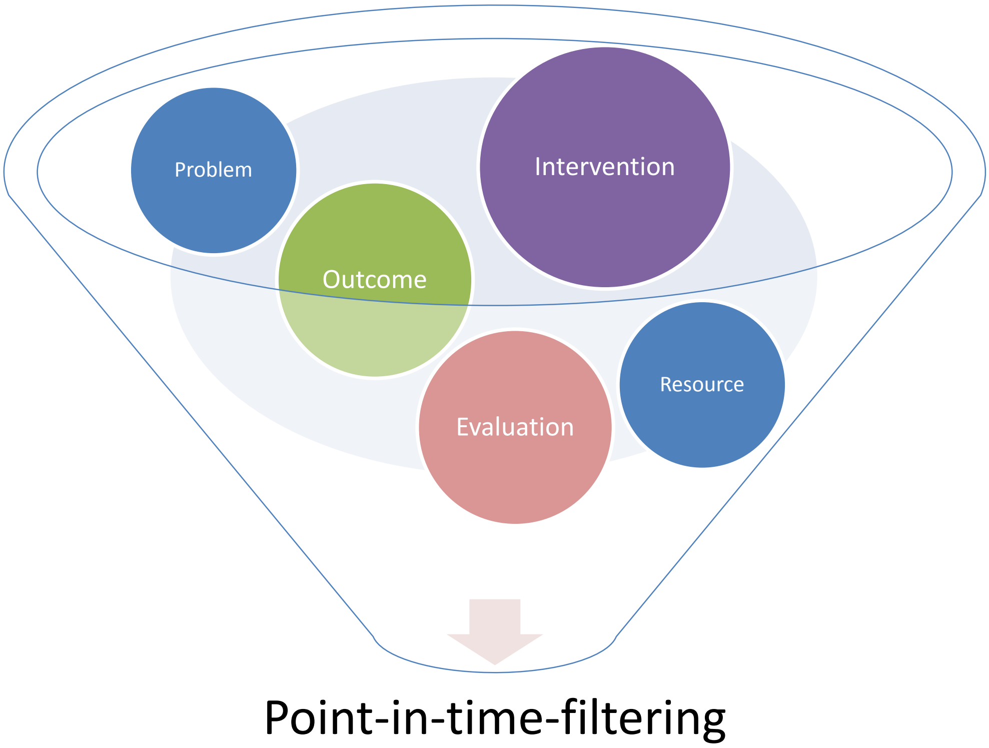


Figure 5.1: Extended OGDL point-in-time-filtering visualisation

In this chapter, we propose a new approach to derive composite data driven clinical success indicators from clinical trial datasets, and compare the results with published indicators from existing clinical guidelines. We propose that data driven clinical risk and related resource and implementation indicators be identified through machine learning of past evidence. Disparate and heterogeneous health data is semantically integrated through use of a novel primitive upper ontology, developed for the purpose of risk management in a problem solution learning framework. We propose the development of a health research collaboration system as depicted to mine and peer-review quantitative and qualitative scientific and observational clinical data, including streaming electronic health records. This will support the development of a best practice clinical practice guideline assessment framework with evidence based on the collaboration platform’s health knowledge repository. The aim is to provide best-practice holistic risk management at the care interface, to help close the gap between the strategic intentions of clinical guidelines and its real-world impact, through an acceleration of holistic collaborative clinical risk management research. We aimed at meeting national and international eHealth goals to support evidence based medicine through standardized and interoperable electronic health records; and the transfer of expert policy recommendations into appropriate composite clinical success indicators for real-time point-of-care clinical risk decision-support. This research has significance for academic and applied researchers, health service planning and management professionals, and health workers, as an introduction to a new approach to collaborative optimised preventive risk management, using a semantic web based collaboration platform for risk management. The approach has wider applicability to public and environmental risk management.

5.2. Relevant Work

5.2.1. Health Data Semantic Interoperability

The European Commission reported in [94] that a wide range of international stakeholders have a growing concern to address the more complex and generalised challenges of patient safety and the cost-effective and equitable use of healthcare resources, though the achievement of full semantic interoperability of health data. The report describes current research efforts such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) and HL7, as 'large and unwieldy' and not helpful in achieving health data semantic interoperability. They stated that there is an urgent need for the development of a framework for aggregation of electronic health data to produce public health indicators and alerts.

5.2.2. Upper Ontology Development

A number of studies have identified that an upper ontology is necessary for accurate data integration in real-time [95-97]. Examples of application specific development and implementation of upper ontology semantic reasoning capacity are found in the domains of military and security risk management, for example in [98]; and of environmental sustainability management, as exemplified by [99]. These applications are in-house decision-support tools, aimed at real-time risk prevention and mitigation; domain knowledge specific as well as local process specific; and do not use a universally acceptable user-friendly primitive upper ontology that would enable its wider semantic reasoning application in terms of risk management. Current approaches to upper ontology development have thousands of elements, including BFO, Cyc, DOLCE, GFO, PROTON, Sowa's ontology and SUMO. They typically have 2,000 to 10,000 elements (classes, relations) with complex interactions among them, and do not share a common approach or method to compare their performance in practice [100].

The SUO WG Suggested Upper Merged Ontology (SUMO) consists of approximately 4,000 assertions (including over 800 rules) and 1,000 concepts [101]. SUMO’s basic entities are physical, including object and process terms, and the abstract ‘thinking’ terms quantity, attribute, set or class, relation, proposition, graph, and graph element. For practical application, SUMO has been found to need extensive addition to its terms for the purposes of risk management team roles and tasks for an in-house application [102].

PSL is an International Standard (ISO 18629) modular, extensible first-order logic ontology [82] that aims to capture upper ontology concepts required for manufacturing and business process specification. PSL does not serve as a standard primitive upper ontology, as it has over 300 concepts across 50 extensions of a common core theory (PSL-Core), each with a set of first-order axioms written in Common Logic (ISO 24707).

The lack of universal agreement about a standard upper ontology with a limited number of primitive elements that would enable user-friendly mapping and merging of all existing domain ontologies in a collective knowledge model constitutes a significant research challenge [100]. As a meta-ontology for inter-ontology mapping its meta-level concepts would support collaborators to accurately map between sets of classes of different ontologies with differences in meaning [104]. Such a standard primitive upper ontology would guide data fusion and development of mathematical algorithms [105] for event risk computation, and enable collaborators to use First-Order Logic for the purpose of collaborative risk management.

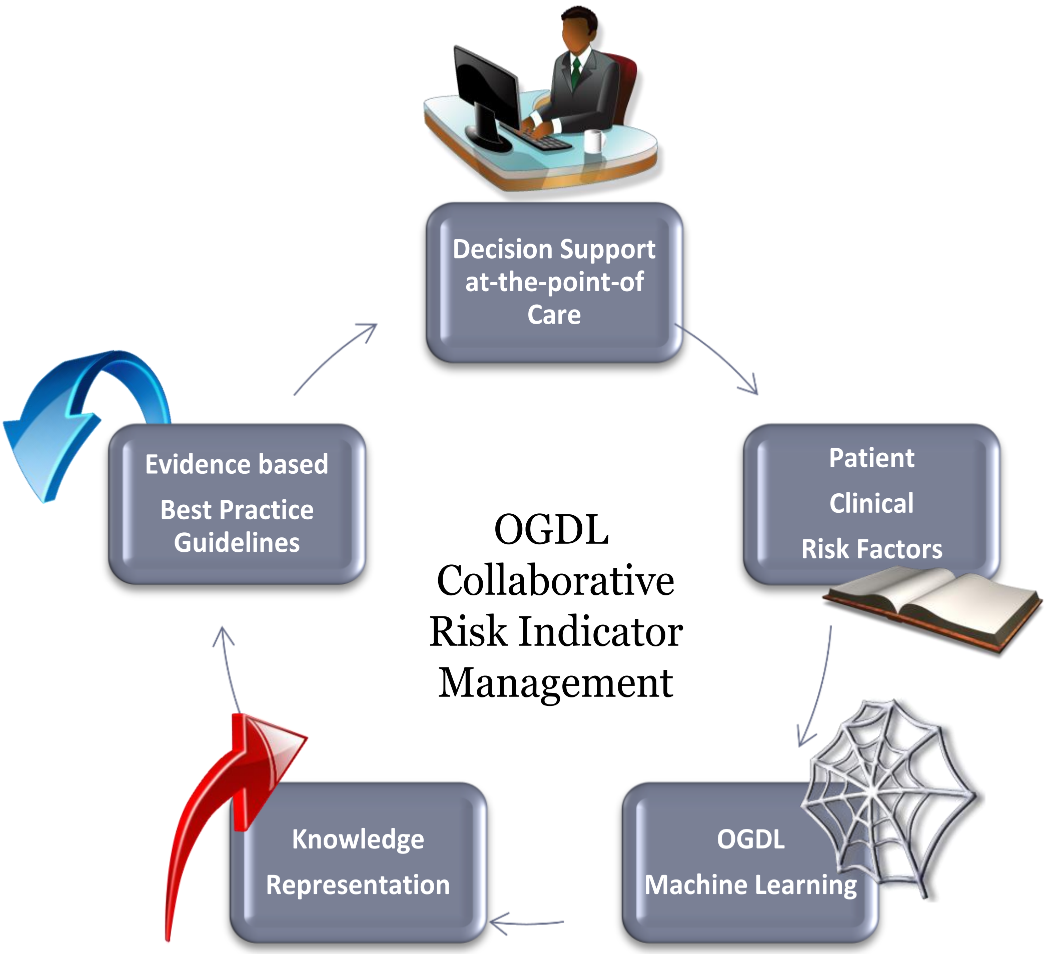

Decision Support at-the-point-of Care
Patient Clinical Risk Factors
OGDL Machine Learning
Knowledge Representation
Evidence based Best Practice Guidelines
OGDL Collaborative Risk Indicator Management

Figure 5.2: The OGDL framework towards Collaborative Risk Indicator Management.

5.3. Extension of OGDL Framework with FLORM

5.3.1. FLORM

In order to support development of a risk knowledge repository; and to enable semantic reasoning to deliberate consensus on improved success and risk indicators, we propose an extension of our OGDL framework [106] through a novel First-Order Logic Primitive (with less than 100 elements) Upper Ontology for Risk Management, (FLORM),. We extend OGDL to extract semantic cluster patterns of past evidence of resource and intervention success for specific problems from the knowledge repository, which is organized through FLORM in a problem-solution framework. This enables machine learning of data driven composite holistic success indicators from the knowledge repository, as an integration of risk indicators with successful resource and intervention indicators. This is significant for an evidence-based approach to risk management, as depicted in Figures 5.1 and 5.2. The proposed primitive upper ontology consists of 4 layers as shown in Table 5.1.

Table 5.1: Proposed FLORM Upper Ontology Layers

Layer 1	5 high-level meta-level concepts (general entities that do not belong to a specific problem domain, and thereby would lead naturally to a categorization scheme for existing thesauri, encyclopedias, indices, etc.). These are: Problem, Resource, Implementation, Outcome and Evaluation. These concepts are perdurants, which are entities that can only be seen partly at any given snapshot in time.
Layer 2	The proposed primitive upper ontology has a mid-level ontology of 6 meta-level concepts: These are: What, Who, Where, When, Why, How. In upper ontologies, these concepts are endurants, which are those entities that can be observed-perceived as a complete concept, at no matter which given snapshot of time.
Layer 3	The 5 low-level meta-level concepts are IF, AND, THEN, ELSE, ELSEIF, which enable advanced querying of mappings to the upper ontology, using SQL. These concepts are endurants.
Layer 4	The lowest-level meta-level concepts define data linkages between unique concepts, and consist of 3 triplets organized in strings. Figure 3 demonstrates the OGDL extension with FLORM in terms of success factor derivation.

The OGDL Composite success indicator analysis algorithm is defined in Algorithm 4 and the prototype is depicted in Figure 5.4 and Figure 5.5. The algorithm firstly maps the semantic clusters of available resources and intervention options specific to a set of problems, and then identifies the evidence-based clusters of resources and interventions that conform to both of this as well as to the user provided set of outcome and evaluation success factors.

Algorithm 4: OGDL Composite Success Indicator Analyser
Success indicator algorithm

Input:

A Set of Clinical PROBLEM, RESOURCE, IMPLEMENTATION, OUTCOME and EVALUATION factors.

Output:

A Set of composite SUCCESS indicators per PROBLEM.

IF (PROBLEM, 5WH) = (x1, x2, xn,…) AND AVAILABLE (RESOURCE, 5WH) = (y1, y2, yn,…) AND OPTIONS (INTERVENTION, 5WH) = (z1, z2, zn,…) AND SUCCESS (OUTCOME, 5WH) $\subset$ (a1,a2,an,…) AND SUCCESS (EVALUATION, 5WH) $\subset$ (b1,b2,bn,…)

THEN = (SUCCESS INDICATOR, 5WH) = (c1,c2,cn,…) WHERE C $\subset$ (x1y1z1a1b1, x1y2z1a1b1, x1y3z1a1b1, …)

An example as applied to our clinical dataset:

```
IF PROBLEM     = WHO      (Adult Age Group, Gender)
                 WHAT     (Age Years, Weight Kg, Ankle/tibia/fibula fracture or dislocation, Average of Minutes Between Procedure And Last Food, Average of Minutes Between Procedure And Last Fluid)
AND  INTERVENTION  =  WHAT   (Sedation Combination contains Propofol IV)
AND  OUTCOME       =  WHAT   (Level of Sedation 5 or 6)
AND  EVALUATION    =  WHAT   (Successful Procedure)
THEN
     RESOURCE      =  WHO    (Staff seniority, Staff discipline, ED doctor status)
AND  INTERVENTION  =  WHEN   (Procedure timing within the sedative drug's pharmaco-logical-ly active time, Minutes Between Procedure And Medication)
                      WHAT ( Sedation Drug Groups used, All Drugs used in Combination)
                      HOW    (Average Drug given mg per kg)
```

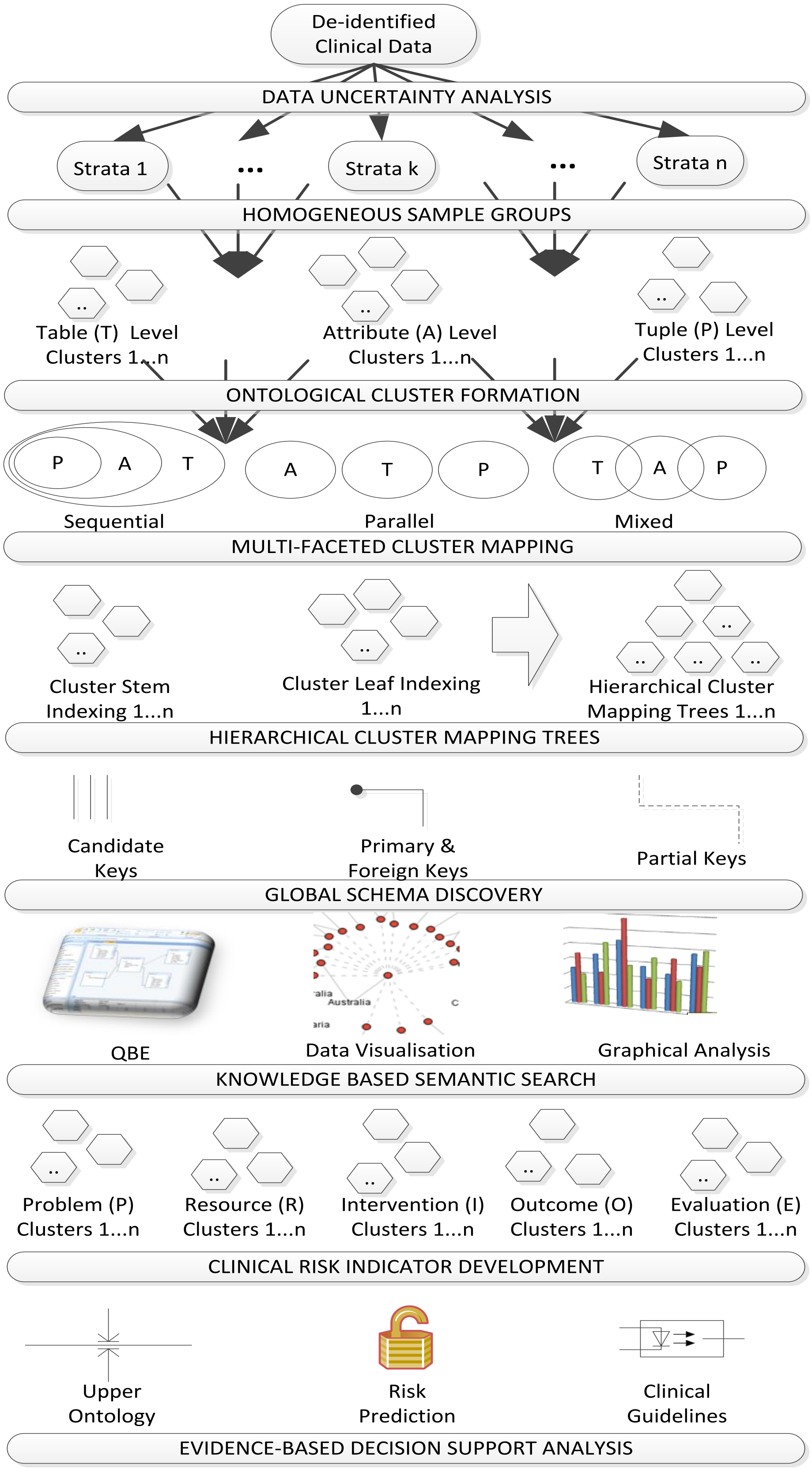

De-identified Clinical Data
DATA UNCERTAINTY ANALYSIS
Strata 1
...
Strata k
...
Strata n
HOMOGENEOUS SAMPLE GROUPS
..
Table (T) Level Clusters 1...n
..
Attribute (A) Level Clusters 1...n
..
Tuple (P) Level Clusters 1...n
ONTOLOGICAL CLUSTER FORMATION
P
A
T
A
T
P
T
A
P
Sequential
Parallel
Mixed
MULTI-FACETED CLUSTER MAPPING
..
Cluster Stem Indexing 1...n
..
Cluster Leaf Indexing 1...n
..
..
..
..
Hierarchical Cluster Mapping Trees 1...n
HIERARCHICAL CLUSTER MAPPING TREES
Candidate Keys
Primary & Foreign Keys
Partial Keys
GLOBAL SCHEMA DISCOVERY
QBE
Australia
Data Visualisation
Graphical Analysis
KNOWLEDGE BASED SEMANTIC SEARCH
..
Problem (P) Clusters 1...n
..
Resource (R) Clusters 1...n
..
Intervention (I) Clusters 1...n
..
Outcome (O) Clusters 1...n
..
Evaluation (E) Clusters 1...n
CLINICAL RISK INDICATOR DEVELOPMENT
Upper Ontology
Risk Prediction
Clinical Guidelines
EVIDENCE-BASED DECISION SUPPORT ANALYSIS

Figure 5.3: The general architecture of extended version of the OGDL framework.

5.4. FLORM Real-World Data Mapping Using Semantic Technologies

We propose a unique approach of triplets that map the upper ontology to existing semantic technologies and to real-world data, including peer-review of evidence and expert opinions in regards to risk management, as discourse threads that can be data mined. Threads are collaborator time-stamped strings of semantic primitive upper ontology use: the ontology performs as a ‘metalanguage’, that relates problem-solution risk management concepts to the use of unique concepts. These concepts are linked in RDF triples, where the RDF vocabulary elements are represented by unique health concepts. Strings of triplets are threaded to link system applications through a semantic web middleware architecture to a collaboration platform for real-time risk management. This approach enables semantic interoperability with semantic technologies, as it extends the semantic web concept of a Resource Description Framework (RDF) graph. An RDF graph typically consists of triplet subject-predicate-object expressions, and we not use triplets stores as unique concept lists, as this format does not enable the flexibility needed to model different choices and their real-world impacts as statistical concordances. Instead we maintain the triplet format of subject-predicate-object as snapshot in time map-able using IF THEN, etc. by users for real-world data linkage. We support collaborators to move away from a freeform sentence format for communication and to conform to contribution to 5WH question and answer option development. This leads to threads of problem related answer (ranked) support as statistical use of unique concepts, mapped to the upper ontology and to the unique concept topical index, in a particular collaboration discourse.

We extend the triplet format to an entity–attribute–value model with an object oriented design; this enables robust data source, level of evidence/agreement/satisfaction referencing as meta-data specification by users. This constitutes a bridging of a current significant gap in research, as multiple triplets of subject-predicate-object AND entity–attribute–value can now be threaded together to enable logical discourse threads that are enriched in multiple ways through metadata links: hyperlinks to the multi-media source data; Wikipedia style URL page definitions of unique concepts referenced to existing health coding indices including SNOMED-CT, HL7 and ICD-10; contributing collaborators’ real-world data including electronic health records; published web table and freeform text web data; to visual depictions of forecasting and consensus approximation; etc., to continue unambiguous communication for accelerated and reliable evidence and consensus development.

Our primitive upper ontology approach enables representation of clinical knowledge in a format that would permit robust knowledge repository additions and accurate knowledge flow for decision-support application. Collaborative risk knowledge repository creation is quality assured through semantic reasoning, in a process of timely evidence-based expert consensus approximation to prevent preventable risk within windows of opportunities, as suggested by machine learning of past evidence. This would support best-practice dissemination for reliable and timely preventive risk management at the care interface, as well as at service funding and planning stages. At the point of care, the use of the primitive upper ontology enables decision-support for specific patient problem's, supported by data mined patterns of past interventions' short-and-long-term costs and benefits for patients with similar complex clinical risk profiles.

5.5. Experimental Evaluation for Clinical Success Factors

5.5.1. Clinical Trial Datasets

We analysed an Australian clinical trial datasets of 2,623 patients that was collected from eleven Australian public hospital emergency departments, between January 2006 and December 2008 [107]. Patients were included if a sedative drug was administered for an emergency department procedure, and data include detailed risk knowledge relating to patient problems; procedural staff; procedural drugs; clinical procedures; patient outcomes and procedural evaluations. We derive data driven composite clinical success factors from successful patient outcomes for the purpose of future evidence-based best-practice decision-support at the point of care, and compare our recommendations to published risk indicator data extracted from the same datasets using the status quo approach to consensus deliberation on clinical research findings for improved health outcomes.

5.5.2. FLORM Mapping to the Datasets

Expert knowledge was used to map the datasets row and column headings to FLORM. Dataset column headings were categorized as 'experimental' vs. 'classification' factors; and were mapped to the primitive upper ontology using the intuition that a patient's demographic and clinical risk factors, including their reason for having to be treated, represent ‘problems’. These factors were therefore mapped to the problem (who, what, where, when, how, and why) FLORM concepts. The experimental factors (that can be changed in future, organizationally) were mapped to FLORM resource (5WH), intervention (5WH), outcome (5WH) and evaluation (5WH). We perform statistical data extraction given a specific problem set, and available knowledge of past successful resource choices and intervention outcomes. We introduce the concept of problem-based semantic reasoning using FLORM, for example:

IF Problem = X AND Outcome = Success AND Evaluation = Success, THEN Resource is approximated to be Y and Intervention is approximated to by Z, given available evidence.

In regards to this dataset, the following mapping was performed to extract the Composite

Clinical Success Indicator:

IF [Risk Indicator = Generic Age Group + Injury Type + Minutes since Last Food/Fluids/Alcohol] AND [Evaluation Indicator (Successful procedure, Level of Sedation 5 or 6)] of composite associated [Risk Indicator + Resource Indicator + Intervention Indicator + Outcome Indicator]

THEN [Resource Indicator = Hospital Care Discipline + ED Doctor Status + Staff Seniority + Drug Administrated and Route] AND [Intervention Indicator = Combination Drug Group + Aver-age Minutes Between Procedure and Medication + Procedure Performed within Pharmacological Active Time + Average Given Dose mg Per kg].

5.5.3. Knowledge Extraction Procedures

We used FLORM first-order logic for semantic reasoning and knowledge extraction to derive Composite Success Indicators from the dataset. We extracted problem, resource and intervention knowledge relating to successful outcomes and evaluations for adult patients that had emergency procedural sedation that included the drug Propofol administered intravenously, alone or with other sedatives. Existing clinical risk findings from the same dataset, performed by the clinical trial researchers, were extracted using the status quo clinical research approach of statistical multivariate logistic regression. Their findings are focused on the occurrence of risk factors in terms of procedural failure. Some of their conclusions were that Propofol had the highest failure rate of all sedative drugs used as a single sedative agent (5.9%, 95% CI 4.6–7.6) [108]; increased body weight and specific procedures, such as hip reduction, were associated with significantly higher failure rates [108]; and that increasing age and level of sedation, pre-medication with fentanyl, and sedation with Propofol, midazolam or fentanyl were risk factors for an airway event (P < 0.05) [80].

Table 5.2: Successful Procedure, Level of sedation 5 or 6, Adult patients, Ankle/tibia/fibula procedures, Propofol IV used alone or in combination with other sedatives

	Age in Years	Weight in Kg	Average of Minutes Between Procedure And Last Food	Average of Minutes Between Procedure And Last Fluid	Average of Minutes Between Procedure And Last Alcohol	Average of Procedure timing within the sedative drug's pharmacologically active time	Average of Minutes Between Procedure And Medication
COUNT (Observations)	67.00	61.00	61.00	61.00	6.00	51.00	64.00
AVG	43.07	81.02	559.38	477.57	-146.00	0.02	0.14
MAX	63.00	170.00	14520.00	14520.00	374.00	1.00	10.00
MIN	17.00	50.00	-1100.00	-1143.00	-1120.00	-1.00	-1.00
STDEV	12.81	21.10	1877.67	1880.18	745.39	0.31	1.26
STDEVP	12.71	20.93	1862.21	1864.71	680.44	0.30	1.25
VAR	164.10	445.35	3525636.47	3535088.72	555606.00	0.09	1.59
VARP	161.65	438.05	3467839.15	3477136.44	463005.00	0.09	1.56

Table 5.3: Propofol IV initial dose in combination with other sedatives

	COUNT (Observations)	AVG	MAX	MIN	STDEV	STDEVP	VAR	VARP
Endone PO, Midazolam IV, Morphine IV, Propofol IV	1.00	170	170	170		0		0
Fentanyl IV, Ketamine IV, Propofol IV	1	10	10	10		0		0
Fentanyl IV, Metoclopramide IV, Midazolam IV, Morphine IV, Propofol IV	1	50	50	50		0		0
Fentanyl IV, Morphine IV, Propofol IV	6	125	300	20	102.52	93.59	10510	8758.33
Fentanyl IV, Nurofen PO, Propofol IV	1	200	200	200		0		0
Fentanyl IV, Panadeine Forte PO, Propofol IV	1	200	200	200		0		0
Fentanyl IV, Propofol IV	20	98.75	280	25	66.92	65.23	4478.62	4254.69
Ketamine IV, Morphine IV, Propofol IV	2	81	90	72	12.73	9	162	81
Metoclopramide IV, Morphine IV, Propofol IV	2	85	120	50	49.50	35	2450	1225
Midazolam IV, Morphine IV, Nitrous oxide % INH , Propofol IV	1	110	110	110		0		0
Midazolam IV, Morphine IV, Propofol IV	2	100	100	100	0	0	0	0
Morphine IV, Nitrous oxide % INH , Propofol IV	1	80	80	80		0		0
Morphine IV, Panadeine Forte PO, Propofol IV	1	120	120	120		0		0
Morphine IV, Propofol IV	17	97.94	200	50	47.34	45.92	2240.81	2109
Propofol IV	13	94.23	175	40	44.62	42.87	1991.03	1837.87
GRAND TOTAL	75	98.49	300	10	58.74	58.34	3449.85	3403.85

Table 5.4: Propofol IV subsequent dose in combination with other sedatives

	Fentanyl IV, Metoclopramide IV, Midazolam IV, Morphine IV, Propofol IV	Fentanyl IV, Midazolam IV, Propofol IV	Fentanyl IV, Morphine IV, Propofol IV	Fentanyl IV, Propofol IV	Metoclopramide IV, Morphine IV, Propofol IV	Morphine IV, Propofol IV	Propofol IV	Grand Total
COUNT (Observations)	1.00	2.00	2.00	17.00	1.00	9.00	5.00	37.00
AVERAGE	40.00	35.00	72.50	45.15	75.00	42.22	47.53	46.36
MAX	40.00	40.00	100.00	100.00	75.00	70.00	100.00	100.00
MIN	40.00	30.00	45.00	20.00	75.00	20.00	11.00	11.00
STDEV		7.07	38.89	18.76		17.16	36.97	22.17
STDEVP	0.00	5.00	27.50	18.20	0.00	16.18	33.07	21.87
VAR		50.00	1512.50	352.10		294.44	1367.09	491.43
VARP	0.00	25.00	756.25	331.39	0.00	261.73	1093.67	478.15

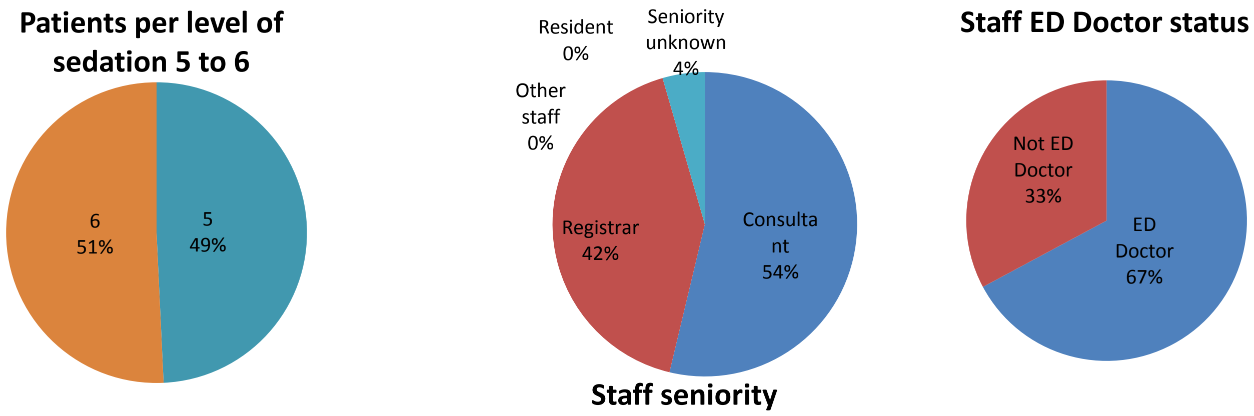


Figure 5.4: From left to right, Patients per level of Sedation 5 or 6, Staff seniority, Staff ED Doctor status.

In comparison to the standard clinical research approach, we mapped the datasets to FLORM categories of problems, resources, interventions, outcomes and evaluations (see Figure 5.5 and Figure 5.6). We then filtered our datasets for successful outcomes and evaluations, and continued to perform OGDL semantic clustering at table, column and tuple levels, as described in [106]. Through this process, we derived semantic clusters of resources and interventions that were associated with success for specific sets of patient problems. Following this process, we derived statistical data for each experimental factor. This demonstrates our novel approach to evidence-based decision-support at the clinical interface; and to evidence-based continuous improvement of machine readable clinical guidelines. Reliability of this approach is dependent on sufficient patient case data being incorporated in the knowledge repository for data driven findings to be of statistical significance; and on integration of weighted expert opinion to validate the levels of evidence and agreement with best-practice.

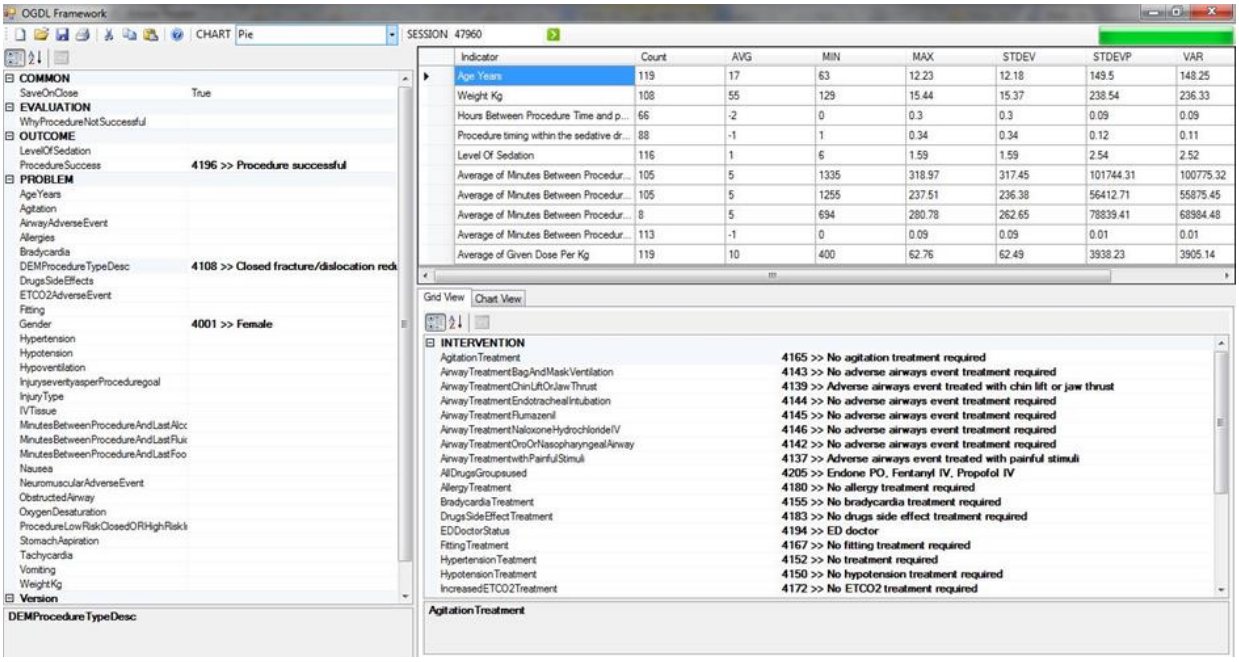


CHART Pie

SESSION 47960

COMMON	
SaveOnClose	True
EVALUATION	
WhyProcedureNotSuccessful	
OUTCOME	
LevelOfSedation	
ProcedureSuccess	**4196 >> Procedure successful**
PROBLEM	
AgeYears	
Agitation	
AirwayAdverseEvent	
Allergies	
Bradycardia	
DEMProcedureTypeDesc	**4108 >> Closed fracture/dislocation redu**
DrugsSideEffects	
ETCO2AdverseEvent	
Fitting	
Gender	**4001 >> Female**
Hypertension	
Hypotension	
Hypoventilation	
InjuryseverityasperProceduregoal	
InjuryType	
IVTissue	
MinutesBetweenProcedureAndLastAlc	
MinutesBetweenProcedureAndLastFlui	
MinutesBetweenProcedureAndLastFoo	
Nausea	
NeuromuscularAdverseEvent	
ObstructedAirway	
OxygenDesaturation	
ProcedureLowRiskClosedORHighRiskI	
StomachAspiration	
Tachycardia	
Vomiting	
WeightKg	
Version	

DEMProcedureTypeDesc

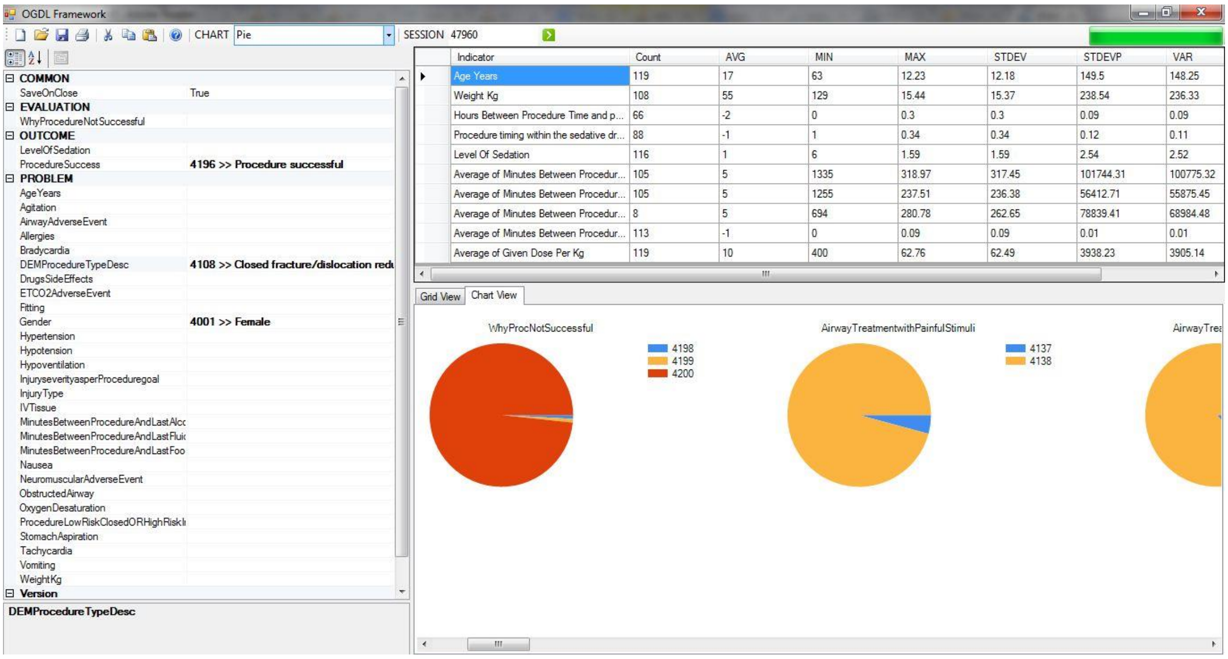
Indicator	Count	AVG	MIN	MAX	STDEV	STDEVP	VAR
Age Years	119	17	63	12.23	12.18	149.5	148.25
Weight Kg	108	55	129	15.44	15.37	238.54	236.33
Hours Between Procedure Time and p...	66	-2	0	0.3	0.3	0.09	0.09
Procedure timing within the sedative dr...	88	-1	1	0.34	0.34	0.12	0.11
Level Of Sedation	116	1	6	1.59	1.59	2.54	2.52
Average of Minutes Between Procedur...	105	5	1335	318.97	317.45	101744.31	100775.32
Average of Minutes Between Procedur...	105	5	1255	237.51	236.38	56412.71	55875.45
Average of Minutes Between Procedur...	8	5	694	280.78	262.65	78839.41	68984.48
Average of Minutes Between Procedur...	113	-1	0	0.09	0.09	0.01	0.01
Average of Given Dose Per Kg	119	10	400	62.76	62.49	3938.23	3905.14

INTERVENTION	
AgitationTreatment	**4165 >> No agitation treatment required**
AirwayTreatmentBagAndMaskVentilation	**4143 >> No adverse airways event treatment required**
AirwayTreatmentChinLiftOrJawThrust	**4139 >> Adverse airways event treated with chin lift or jaw thrust**
AirwayTreatmentEndotrachealIntubation	**4144 >> No adverse airways event treatment required**
AirwayTreatmentFlumazenil	**4145 >> No adverse airways event treatment required**
AirwayTreatmentNaloxoneHydrochlorideIV	**4146 >> No adverse airways event treatment required**
AirwayTreatmentOroOrNasopharyngealAirway	**4142 >> No adverse airways event treatment required**
AirwayTreatmentwithPainfulStimuli	**4137 >> Adverse airways event treated with painful stimuli**
AllDrugsGroupsused	**4205 >> Endone PO, Fentanyl IV, Propofol IV**
AllergyTreatment	**4180 >> No allergy treatment required**
BradycardiaTreatment	**4155 >> No bradycardia treatment required**
DrugsSideEffectTreatment	**4183 >> No drugs side effect treatment required**
EDDoctorStatus	**4194 >> ED doctor**
FittingTreatment	**4167 >> No fitting treatment required**
HypertensionTeatment	**4152 >> No treatment required**
HypotensionTreatment	**4150 >> No hypotension treatment required**
IncreasedETCO2Treatment	**4172 >> No ETCO2 treatment required**

AgitationTreatment

Figure 5.5: Extended OGDL Framework Prototype for FLORM with Grid View

Table 5.2 shows the Successful Procedure, Level of sedation 5 or 6, Adult patients, Ankle/tibia/fibula procedures, Propofol IV used alone or in combination with other sedatives. From the results, it is significant that the average age of adults that were associated with success was 43 years; this corresponds to the heuristic that youth impacts positively on health outcomes. Similarly, average weight for the group was 80kg, which also corresponds to the heuristic of the impact of a healthy weight on successful outcomes. It is important to note that the average minutes between food and fluid consumption that was associated with success was more than 4 hours - this confirms the belief that these factors have a significant impact on the success of outcomes. This observation is an example of how 'anecdotal' patient case data can help prioritise future research efforts to actively reduce poor outcomes. Also significant was the finding that the administration of the sedative drug needed to be within the time period that would assure its pharmacological activity during the procedure.



OGDL Framework

CHART Pie

SESSION 47960

COMMON	
SaveOnClose	True
EVALUATION	
WhyProcedureNotSuccessful	
OUTCOME	
LevelOfSedation	
ProcedureSuccess	**4196 >> Procedure successful**
PROBLEM	
AgeYears	
Agitation	
AirwayAdverseEvent	
Allergies	
Bradycardia	
DEMProcedureTypeDesc	**4108 >> Closed fracture/dislocation redu**
DrugsSideEffects	
ETCO2AdverseEvent	
Fitting	
Gender	**4001 >> Female**
Hypertension	
Hypotension	
Hypoventilation	
InjuryseverityasperProceduregoal	
InjuryType	
IVTissue	
MinutesBetweenProcedureAndLastAlcc	
MinutesBetweenProcedureAndLastFluic	
MinutesBetweenProcedureAndLastFoo	
Nausea	
NeuromuscularAdverseEvent	
ObstructedAirway	
OxygenDesaturation	
ProcedureLowRiskClosedORHighRiskI	
StomachAspiration	
Tachycardia	
Vomiting	
WeightKg	
Version	

DEMProcedureTypeDesc

Indicator	Count	AVG	MIN	MAX	STDEV	STDEVP	VAR
Age Years	119	17	63	12.23	12.18	149.5	148.25
Weight Kg	108	55	129	15.44	15.37	238.54	236.33
Hours Between Procedure Time and p...	66	-2	0	0.3	0.3	0.09	0.09
Procedure timing within the sedative dr...	88	-1	1	0.34	0.34	0.12	0.11
Level Of Sedation	116	1	6	1.59	1.59	2.54	2.52
Average of Minutes Between Procedur...	105	5	1335	318.97	317.45	101744.31	100775.32
Average of Minutes Between Procedur...	105	5	1255	237.51	236.38	56412.71	55875.45
Average of Minutes Between Procedur...	8	5	694	280.78	262.65	78839.41	68984.48
Average of Minutes Between Procedur...	113	-1	0	0.09	0.09	0.01	0.01
Average of Given Dose Per Kg	119	10	400	62.76	62.49	3938.23	3905.14

Figure 5.6: Extended OGDL Framework Prototype for FLORM with Chart View

Table 5.3 shows the Propofol IV initial dose in combination with other sedatives. It is significant to note that there is a large variation in the average dose of Propofol that was administered intravenously in combination with other sedatives, with associated sedative success. From Table 5.4, we observe that the subsequent doses of intravenous Propofol that was administered with success in combination with other sedatives, after the initial (loading) dose, were significantly less than the initial dose, and again varied significantly in combination with different sedatives.

Figure 5.4 shows a set of composite success indicator factors. Clinicians evaluated the patient’s level of sedation on a scale of 1 (lowest) to 6, and we defined 'successful sedation' as a level of 5 or 6. It is significant to note that for ankle, tibia and fibula procedures, success was associated with senior staff with experience, and not with junior staff members. Two thirds of successful procedures were associated with doctors with experience in the emergency department, as opposed to doctors that were from other hospital departments and performed the procedure.

5.6. Summary

In this chapter, we have addressed current research challenges for semantic knowledge management. [104] describes T as a first-order theory, a set of first-order sentences closed under logical entailment and describes the challenge to find new core theories in the case when a reduction of a new theory T, does not exist, but T is reducible to an extension of the core hierarchies. In this chapter, we introduced a unique root theory as a primitive upper ontology, which is a meta-language across ontology repositories. We address the challenge outlined in [104] by proposing that data-driven new knowledge of success indicators are mapped to the upper ontology in a topical problem-solution core hierarchy of unique concept confirmation by collaborators, validating the data driven indicators for real-world and real-time use.

The challenge to incorporate techniques for ontology verification to characterize the models of new theories that are under probation as a result of the decomposing procedure as outlined in [104] is addressed by our proposal for collaborator polling on levels of agreement with theories under probation, as best-practice resources and interventions for problem/evaluation success. We propose to address the challenge through development of a collaboration platform for this purpose, with transparent time-stamped collaborator new theory validation discourse.

The challenge to explore techniques that use the reductions and profiles of theories to generate semantic mappings between ontologies as outlined in [104] is addressed by the proposal for mappings to synonyms of the core unique concepts of FLORM, e.g. problem - issue; challenge; incident. In future work, we aim to introduce the concept of robust updates of unique concepts in Wikipedia pages, as aggregated consensus on its meaning, through a process of transparent peer-review of theories regarding a unique concept's positioning in problems, resources, interventions, etc.

We addressed the research challenge outlined in [104] regarding trunk theories being used to design new extensions of existing ontologies that are reducible to the core hierarchies, given that in any hierarchy, the complete set of trunk theories for a core hierarchy corresponds to the axioms of all complete extensions of the root theory. We enable this through a knowledge repository development approach, where the dataset for a specific problem, and its associated evidence of applications for successful outcomes/evaluations, in terms of resources and its implementations, is used dynamically to derive best-practice composite success factors for specific risk factor sets at a snapshot in time. We proposed that new knowledge be collaboratively developed regarding new unique concepts, and also regarding the use of unique concepts in PRIOE theories.

To summarise, we introduced a novel primitive upper ontology FLORM for risk management, and demonstrated its usefulness in mapping existing health data to a problem-solution framework, to derive composite holistic success factors. We demonstrated how this approach, enhanced by our OGDL framework, enables the development of evidence-based indicators that are both machine readable and human understandable, and an improvement on risk indicators derived on the same dataset using the status quo approach. We discussed FLORM's ability to enable semantic interoperability, which supports a global demand for accurate and reliable semantic integration of disparate and heterogeneous data. We proposed the use of this approach in eHealth to meet the current urgent demand to improve preventive clinical risk management patient outcomes, and equitable service planning and provision.

In the next chapter, as part of our future work, we will introduce the development of a semantic web application for risk management, (SWARM). It will be based on explicit and altruistic public risk-minimisation values and goals, and act as a transparent compliance system to current best practice; as well as a collaborative research platform for improved best practice. We will integrate our proposed risk management problem-solution primitive upper ontology for cross-domain semantic integration of historic risk management activities and outcomes in a shared knowledge base and for semantic reasoning in a conversational expert system approach. The aim will to be enhancing weighted expert opinions using the holistic of teleonic principles [78] for real-world simulation, agile methods, semantic clustering and statistical inference. This will support the alignment of strategy with real-world findings through forward and backward feeding loops for evidence-based holistic data driven composite success indicator development, with minimisation of assumptions and bias.

Through our research contributions, collaborators can be supported to transparently approximate evidence-based consensus on transferable learning of past relevant successful risk interventions. In a real-world, real-time decision-support application, collaborators can provide evidence-based decision-support tailored to a specific local current risk level, the acceptable risk level, the availability of best-practice resources, do-able interventions, timelines, roles and tasks, as reliable holistic success indicators. We aim to introduce a community cloud based collaboration platform which will support collaborators to quality assure accurate data linkage and semantic integration in a shared risk knowledge repository; and to deliberate on best-practice success factors, with support from OGDL and FLORM.

Chapter 6
Conclusions

In this chapter we review the whole thesis and give detailed prospects for future works on our research. First, a short summary is presented as an overview of research with an emphasis on contributions made in Data Linkage and its applications in Clinical Risk Management. We show that our proposed solutions are highly efficient when applied on heterogeneous databases and can be easily adaptable well in practice. Then, future research proposals are introduced. These include the follow-up sub-topics for the future directions presented in this thesis.

6.1. Summary of Results

In this doctorial research, we have provided three significant contributions that can solve the problem of performing Data Linkage through structural analysis on knowledge domain based on probabilistic matching technique. Each of these contributions are summarised below.

First, to deal with the problem of matching pairs, we evaluated and showed how h-gram technique can be effectively and efficiently be used for matching similar records and emphasised its relevance in concept clustering and cluster matching on a probabilistic basis. Through experimental results, we showed that the h-gram record matching is highly significant and advances set-of-sets technique [8] by extending the features of scale based hashing and n-gram techniques.

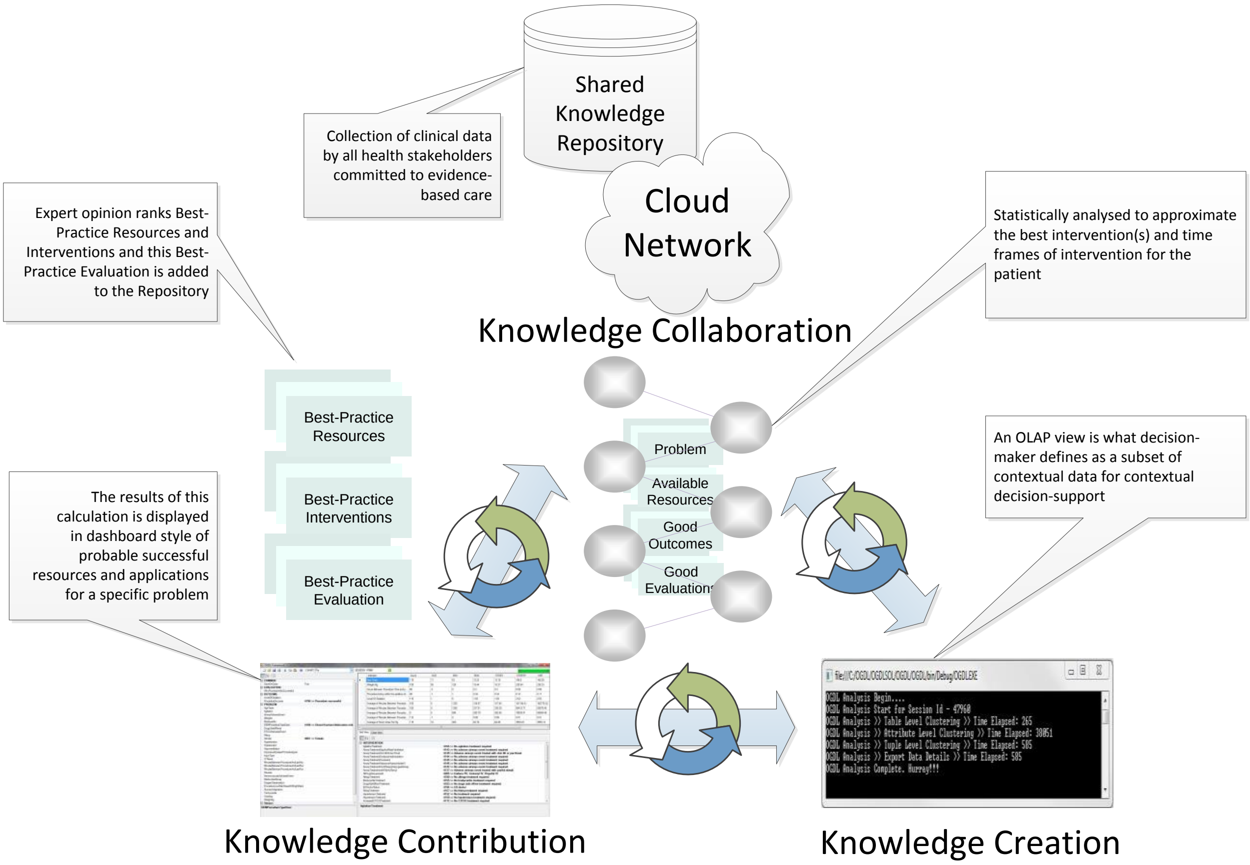
We then provided a highly effective and efficient OGDL framework for querying and integrating heterogeneous databases in the presence of data uncertainties, demonstrating an effective method for identifying how different sets of tables, attributes and tuples can be linked with the primary aim to understand the past and predict the future. To deal with the problem of data uncertainties, we took modular neural-network to the next level through the formal introduction of ranking and classifying ontological characteristics in multiple modules. We figured applying a combination of various methods for solving data linkage problem that is applicable in solving our unique research problem and introduced three unique industry level applications (see section 4.7) which are: 1) we developed the core ‘OGDL Data Miner’ which performs the bulk of our proposed framework tasks at different stages; 2) we developed the ‘OGDL Cluster Search Engine’, an interactive and user-friendly tool to visualize the cluster stem-and-leaves formed across multi-domain data-bases. By clicking on any searchable cluster, the user can drill into its correlated clusters for knowledge discovery and for exploration of ‘chains-of-facts’; 3) we developed the ‘OGDL Performance Monitor’ tool to analyze the scalability of our framework while running on different machines. Through accuracy, performance and scalability experimental tests, we proved that the OGDL approach achieves high quality results and that the development of our framework is highly significant, and important as a step towards advancing the data linkage process.

Finally, we extended OGDL framework and introduced FLORM towards collaborative clinical risk management. FLORM enables the development of machine understandable risk policies, guidelines and standard operating procedures to be developed as shared intelligence, enhanced by peer-review. FLORM aims to support this process through its capacity for first-order logic, by enabling semantic reasoning about the validity and feasibility of various solution options. OGDL supports the process by giving collaborators access to relevant extracted knowledge during consensus deliberation. Collaborator dynamic data linking will be able to support preventive risk management, and related resource and task coordination through access to best-practice success factors at the point-of-decision-making.

In addition, in this concluding chapter, we will introduce the development of cloud based health research collaboration architecture towards the development and dissemination of best-practice clinical guidelines that are reliable, user-friendly, dynamic, tailor-able and timely machine-and-human readable success indicators and early warning risk indicators.

6.2. Future Work

There is a significant problem with the high level of occurrence of preventable morbidity and mortality in healthcare, which directly impacts health service sustainability. Published clinical guidelines are the status quo approach to address this issue, with a variable level of scientific evidence and currency; and they are usually not available in a user-friendly format at the care interface. Our investigation determined that there is an urgent need for user-friendly clinical guidelines that support dynamic evidence-based and patient tailored decision-support in real-time at the point-of-care, and that the current approach to clinical guideline development and dissemination does not support this need.


Shared Knowledge Repository
Collection of clinical data by all health stakeholders committed to evidence-based care
Cloud Network
Expert opinion ranks Best-Practice Resources and Interventions and this Best-Practice Evaluation is added to the Repository
Statistically analysed to approximate the best intervention(s) and time frames of intervention for the patient
Knowledge Collaboration
Best-Practice Resources
Best-Practice Interventions
Best-Practice Evaluation
Problem
Available Resources
Good Outcomes
Good Evaluations
An OLAP view is what decision-maker defines as a subset of contextual data for contextual decision-support
The results of this calculation is displayed in dashboard style of probable successful resources and applications for a specific problem
file:///C:/OGDL/OGDLSOL/OGDL/OGDL/bin/Debug/OGDL.EXE
Knowledge Contribution
Knowledge Creation

```
OGDL Analysis Begin....
OGDL Analysis Start for Session Id - 47960
OGDL Analysis >> Table Level Clustering >> Time Elapsed: 265
OGDL Analysis >> Attribute Level Clustering >> Time Elapsed: 38051
OGDL Analysis >> Tuple Level Clustering >> Time Elapsed: 585
OGDL Analysis >> Export Data Details >> Time Elapsed: 585
OGDL Analysis Complete. Hurray!!!
```

Figure 6.1: Cloud based ontology-guided data integration, knowledge collaboration, contribution and creation

As part of our future directions, we propose a cloud based clinical collaboration for development of a shared health knowledge repository for the purpose of best-practice clinical risk and success indicator development and dissemination as machine and human readable clinical guidelines. We recommend a novel Primitive Upper Ontology for this purpose, and aim to demonstrate its use for integration and machine learning of disparate heterogeneous data and for enhancement of the resulting knowledge by expert opinion to develop consensus on best-practice clinical guidelines. We focus on cloud network architecture, supported by an ontology guided data linkage framework and problem-solution framework, utilizing the primitive upper ontology, for cloud based clinical collaboration. The aim is to investigate and overcome the problems posed by status quo clinical guideline development and dissemination for robust point-of-care decision-support and collaborative research approach which has significant potential to improve health outcomes and related services and its sustainability.

When attempting to develop evidence for best-practice clinical guidelines, such linking has to be accurate and user-friendly to support reliable decision-support, and has to incorporate expert consensus on best-practice [83, 86, 87, 89, 92, and 93]. The status quo of clinical guideline development and dissemination involves both disparate hardcopy and digital resources, in various formats, including drug administration handbooks, protocols, and standards. These resources are not usually user-friendly, have variable levels of accessibility, evidence and currency, and can usually not be tailored to specific patient problems at the decision-making care interface [86, 89].

In this final chapter, as part of our future work, we propose a limited primitive upper ontology for clinical guideline development and dissemination, and introduce the concept of transparent aggregated expert evaluation to examine the wisdom for solutions, based on knowledge of the past, utilizing the upper ontology. We introduce our proposed ontology guided data linkage framework in terms of this approach, and propose its extension through the primitive upper ontology in a semantic cloud based architecture for clinical collaboration. We aim to demonstrate our proposed approach towards the development and dissemination of best-practice clinical guidelines that are reliable, user-friendly, dynamic, tailor-able and timely machine-and-human readable success indicators and early warning risk indicators. We expect this research advancement can demonstrate an improvement in regards to status quo clinical guideline reliability and user-friendliness, highlighting the significance of our work and its future direction in terms of collaborative best-practice clinical guideline development and dissemination.

The rest of this chapter is organized as follows: In Section 6.3 we demonstrate a motivating example and provide our identified recommendations in clinical risk management; in Section 6.4 we propose our future semantic cloud based approach; and in Section 6.5 we draw summary and conclusions of our research work.

6.3. Identified Recommendations

In [102], as a cutting edge approach to collaborative non-profit public service management, the ‘working ontology’’ was defined as in Figure 6.2. The EHMP ontology [99] has many thousands of potential upper level ontology terms, as does other current approaches to developing upper ontologies, despite being a ‘cutting-edge’ approach to preventing preventable public (non-profit) risk. In Australia, as part of our research, we have identified the core recommendations for improved clinical risk management, transferable to other risk management arenas. These recommendations are as follows.

The data linkage process should support all four processes of the knowledge management cycle: Knowledge creation; knowledge structuring; knowledge dissemination; and knowledge application [83]. We acknowledge that accurate data integration in a shared healthcare knowledge repository, as a logical enterprise knowledge warehouse (EKW) that incorporates clinical, administrative, and financial processes [83], will support health care providers to effectively and efficiently assess, control and communicate clinical risk to minimize (unreasonable) clinical risk, and thus support the prevention of medical negligence and/or human error [93].

Integrated (time series) healthcare data, using a clinical risk factor(s) analysis approach, should provide service providers with the necessary capacity for evidence-based healthcare, through the formulation and amendment of clinical guidelines (CPG)s as part of integrated care pathways (ICP)s [87]. ICPs are structured multidisciplinary patient care plans, with detail of the essential steps in the care of patients with a specific clinical risk profile, and integrate preventive functions for medical negligence. This enables any deviation from CPGs to be documented as statistical variance. Such analysis of deviations and variances provide a means for continued systematic audit of clinical practice [86]; improved patient outcomes; and controlled health expenditure [92].

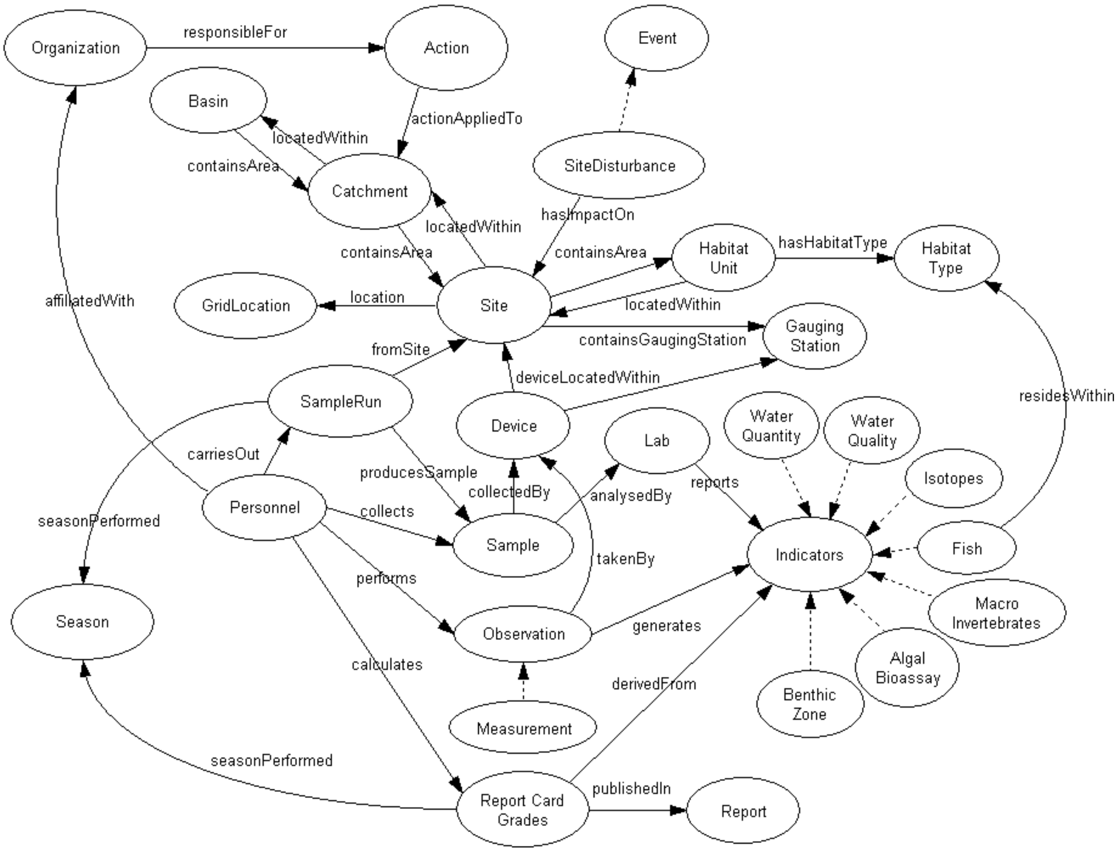


Figure 6.2: Example: Ecosystem Health Monitoring Program (EHMP) Ontology [99]

An industry-wide standard for patient electronic healthcare records (EHCR), linked to clinical guidelines and protocols, is a significant and current research challenge, to ensure best practice clinical risk management. The development of such a 'Gold Standard' presents a globally significant research proposal [83].

An ontology-based knowledge management system needs to integrate domain ontologies of a wide range of data types from disparate data sources to: support data linkage; data integration; semantic analysis; relate management actions to quality indicators for specific entities, regions and periods; identify which actions are having an impact on which parameters using First-Order Logic (enabled by the merged Upper Ontology approach); and adapt related management strategies accordingly [94].

6.4. Clinical Upper Ontology Cloud Framework

In the future work, we aim to optimize prevention of preventable clinical risk; develop quality healthcare resources; distribute limited healthcare resources equitably; identify priority healthcare research needs; and support sustainable healthcare public services. We focus on continuous clinical guideline improvement and user-friendly timely and reliable dissemination as a good example to demonstrate the use of a novel primitive upper to achieve these goals. Besides, it will be interesting to see how related human and machine readable holistic success indicators can be developed and disseminated in a timely manner, collaboratively, using cloud based ontology-guided data integration and expert input.

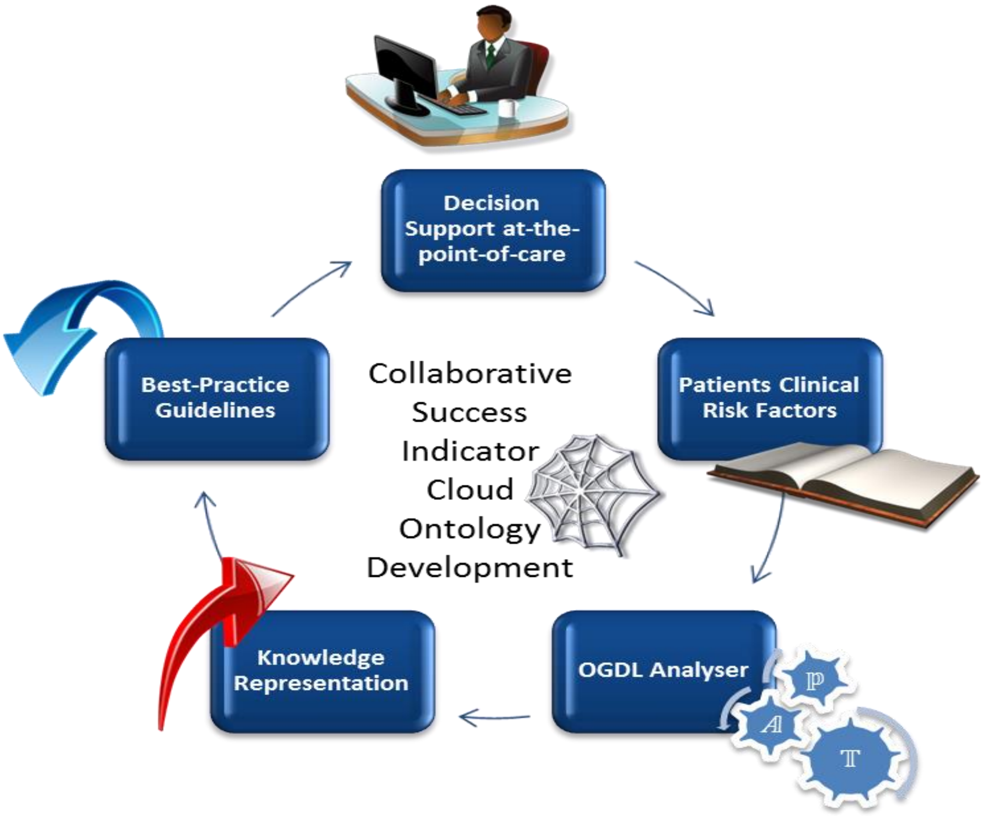

Decision Support at-the-point-of-care
Patients Clinical Risk Factors
OGDL Analyser
P
A
T
Knowledge Representation
Best-Practice Guidelines
Collaborative Success Indicator Cloud Ontology Development

Figure 6.3: An overview of clinical upper ontology cloud framework development.

We plan to propose the concept of an efficient clinical upper ontology cloud framework for collaborative best-practice success indicator development and dissemination. We plan four levels of upper ontology for success indicator development as new proposals of primitive sentences to a novel formal language for collaborative risk management. We propose its continuous improvement according to continuous collaborative expert proposals for changed or new primitive sentences of the formal language that represent holistic composite success indicators. Our proposed rules of proof for the uniqueness of a new success indicator (primitive sentence) relates to a unique mapping of the upper four layers of the ontology, as a new constant of the language, associated by an approximated level of expert consensus within the window of opportunity available to prevent or mitigate risk. Our proposal for a new clinical success indicator, as a new primitive sentence in terms of a successful outcome and evaluation, is mapped by the proposer to the top layers of the upper ontology through use of the third level of the upper ontology (IF, AND, etc.) and the fourth level as question answer options in terms of quantified unique concepts. This corresponds in predicate calculus to the development of a new function, mapping one or more elements in a set (the domain of the function) into a unique element of another set (the range of the function). Elements of the domain and range are objects in the world of discourse. See Figure 6.3 and Figure 6.4 for a depiction of this concept.

A proposed success indicator (primitive sentence) has ‘truth bearing’ criteria for being accepted (and thus asserted) or eliminated, based on an aggregated consensus from all collaborators in a cloud based ontology-guided data integration, at a point-in-time, as levels of agreement and satisfaction, etc. In predicate calculus this corresponds to success indicators as variable symbols. Eliminated proposals for success indicators (new primitive sentences) remain accessible to collaborators for continued deliberation regarding its ‘truth’ or ‘provability’. As new real-world evidence emerges, it is possible for a previously ‘asserted’ success indicator to become less certain, or even eliminated – such change over time is time-stamped per aggregated collaborator input, and thus mappable to the evidence and opinions at the time, in context of the healthcare problems, goals, clinical strategies, standard operating procedures and guidelines of that time. This approach to comparative linguistics supports the process of exploring patterns of inference, where the rules of inference are meaning-constituting, and thus leads to logical harmony of the proposed formal language.

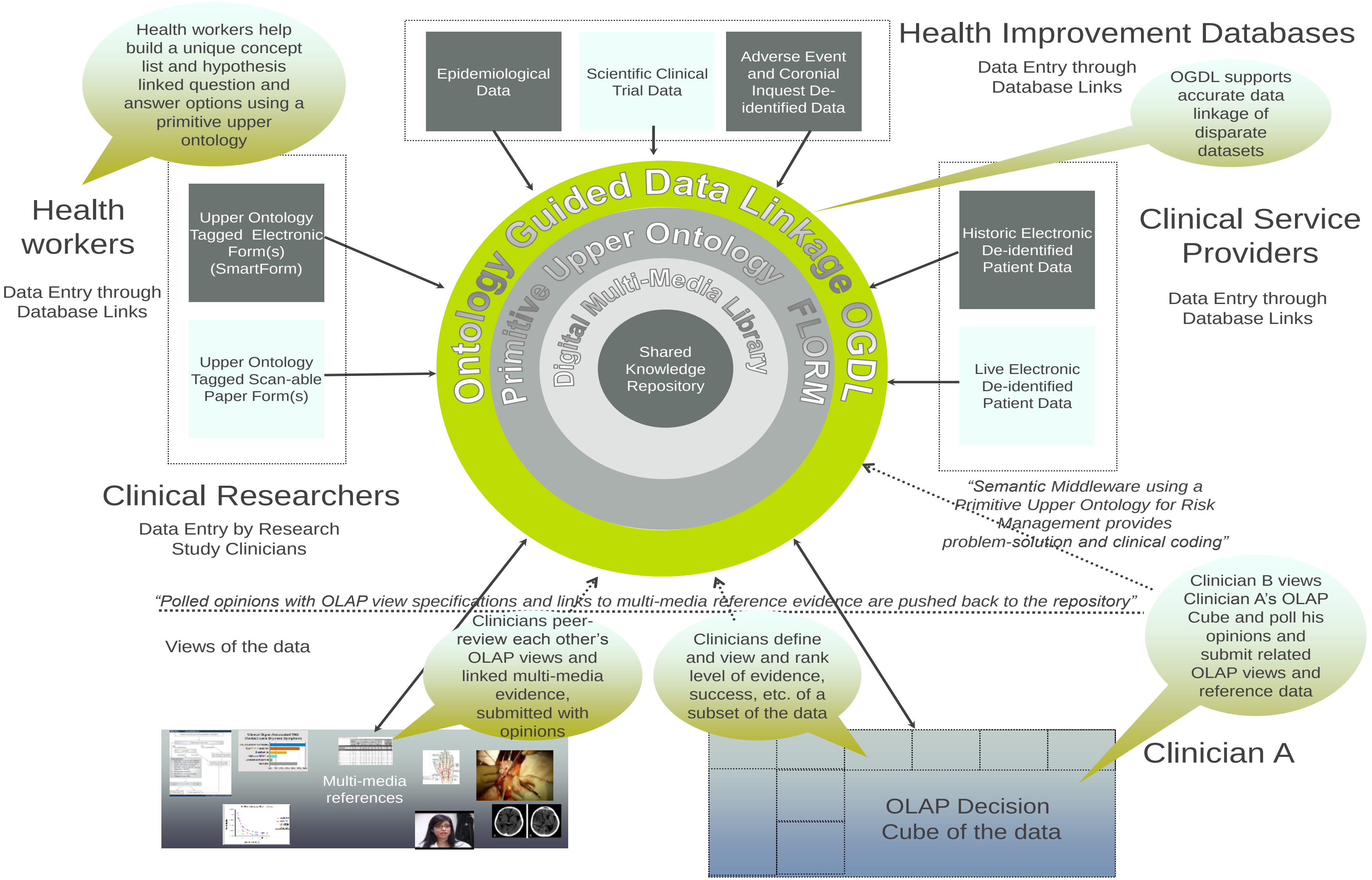

Health workers help build a unique concept list and hypothesis linked question and answer options using a primitive upper ontology
Health workers
Data Entry through Database Links
Upper Ontology Tagged Electronic Form(s) (SmartForm)
Upper Ontology Tagged Scan-able Paper Form(s)
Epidemiological Data
Scientific Clinical Trial Data
Adverse Event and Coronial Inquest De-identified Data
Health Improvement Databases
Data Entry through Database Links
OGDL supports accurate data linkage of disparate datasets
Ontology Guided Data Linkage OGDL
Primitive Upper Ontology FLORM
Digital Multi-Media Library
Shared Knowledge Repository
Historic Electronic De-identified Patient Data
Live Electronic De-identified Patient Data
Clinical Service Providers
Data Entry through Database Links
Clinical Researchers
Data Entry by Research Study Clinicians
"Semantic Middleware using a Primitive Upper Ontology for Risk Management provides problem-solution and clinical coding"
"Polled opinions with OLAP view specifications and links to multi-media reference evidence are pushed back to the repository"
Views of the data
Clinicians peer-review each other's OLAP views and linked multi-media evidence, submitted with opinions
Clinicians define and view and rank level of evidence, success, etc. of a subset of the data
Clinician B views Clinician A's OLAP Cube and poll his opinions and submit related OLAP views and reference data
Multi-media references
OLAP Decision Cube of the data
Clinician A

Figure 6.4: Future directions towards clinical upper ontology cloud framework

6.5. Epilogue

In this chapter, we summarized several directions for our research work. Our proposed future directions towards collaborative development of new success indicators as primitive sentences, constituted of primitive upper ontology variable elements associated with quantifier's, is supported by first-order predicate logic. Temporal analysis supports a collaborative analysis of the different clinical 'proto-languages', based on our proposed formal language for clinical information integration, historically associated with clinical problems, in context of the social changes that led to the documented changes of these proto-languages, in terms of healthcare beliefs and visions, related missions, strategies, protocols, standards, resource and intervention choices, and related historical outcomes and evaluations. We believe this will introduce the concept of transparent aggregated clinical expert evaluation to examine the wisdom for the current best-practice options, based on knowledge of the past, in a process of reflective disclosure, using critical theory. We recommend the use of such transparency of collaborator ranking of level of agreement with risk, quality, safety, success and sustainability of each aspect of a propose clinical success indicator (as a transferable problem-solution approach) to dynamically support local decision-making, given the local problem context and available resources.

In the future, we aim to introduce the concept of collaborative question and answer option development in the cloud, using the primitive upper ontology to develop independent (generic) question and answer options as clinical research hypotheses (success indicators), where the answer option includes quantifiers, for instance the number of staff or time interval between tasks. Thus, the upper ontology continues to evolve, as new questions and answer options are uniquely established and mapped to who, what, where, when, where, how, and IF, AND, THEN, ELSE, ELSEIF of contextual problems, resources, interventions, outcomes and evaluations. It is also very interesting to investigate how this semantic web approach will support collaborators across different sectors, objectives, geographical and resource constraints to improve the efficiency and efficacy of their current services based on relevant evidence of success elsewhere, now and in the past, and accelerate the achievement of their goals.

References

[1] M. Gollapalli, X. Li, I. Wood, G. Governatori, Ontology Guided Data Linkage Framework for Discovering Meaningful Data Facts, *in Proceeding of the 7th International Conference on Advanced Data Mining and Applications (ADMA)*, Beijing, China, 2011, pp. 252-265.

[2] J. Euzenat, P. Shvaiko, Ontology Matching, 1st ed., Berlin Heidelberg: Springer, New York, 2007.

[3] M. Franklin, A. Halevy, D. Maier, From Databases to Dataspaces: a new abstraction for information management, *in J. ACM Special Interest Group on Management of Data (SIGMOD)*, Maryland, Record 34 Issue 4, 2005, pp. 27-33.

[4] S. Fenz, An ontology-based approach for constructing Bayesian networks, *in J. Data & Knowledge Engineering (DKE)*, vol. 73, no. Elsevier, March 2012.

[5] W. Wu, B. Reinwald, Y. Sismanis, R. Manjrekar, Discovering Topical Structures of Databases, *in Proceedings of the 2008 ACM International Conference on Special Interest Group on Management of Data (SIGMOD)*, Vancouver, Canada, 2008, pp. 1019-1030.

[6] E Simperl, Reusing ontologies on the Semantic Web: A feasibility study, *in J. Data & Knowledge Engineering (DKE)*, vol. 68, no. 10, pp. 905-925, Oct 2009.

[7] P. Christen, Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification, *in Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, Nevada, USA, 2008, pp. 151-159.

[8] H. Koehler, X. Zhou, S. Sadiq, Y. Shu, K. Taylor, Sampling Dirty Data for Matching Attributes, *in Proceedings of the 2010 ACM Special Interest Group on Management of Data (SIGMOD)*, Indianapolis, USA, 2010, pp. 63-74.

[9] C. Lee, Automated ontology construction for unstructured text documents, *in J. Data and Knowledge Engineering (DKE)*, vol. 60, no. 3, pp. 547-566, March 2007.

[10] I. Bhattacharya, L. Getoor, Iterative record linkage for cleaning and integration, *in Proceedings of the ACM Special Interest Group on Management of Data (SIGMOD) Workshop on Research issues in data mining and knowledge discovery*, Paris, France, 2004, pp. 11-18.

[11] H. Kim, D. Lee, Parallel Linkage, *in Proceedings of the ACM 16th International Conference on Information and Knowledge Management (CIKM),* Lisbon, Protugal, 2007, pp. 283-292.

[12] Y. Hong, T. Yang, J. Kang, D. Lee, Record Linkage as DNA Sequence Alignment Problem, *in Proceedings of the 6th International Conference on Very Large Data Base (VLDB)*, Auckland, New Zealand, 2008, pp. 13-22.

[13] M. Gagnon, Ontology-based Integration of Data Sources, *in Proceedings of the IEEE 10th International Conference on Information Fusion*, Que, Canada, 2007, pp. 1-8.

[14] A. Bonifati, G. Mecca, A. Pappalardo, S. Raunich, G. Summa, Schema Mapping Verification: The Spicy Way, *in Proceedings of the 11th Internation Conference on Extending Database Technology (EDBT),* Nantes, France, 2008, pp. 1289-1293.

[15] A. Radwan, L. Popa, I.R. Stanoi, A. Younis, Top-K Generation of Integrated Schemas Based on Directed and Weighted Correspondences, *in Proceedings of the 35th International Conference on Management of Data (SIGMOD),* Rhode Island, USA, 2009, pp. 641-654.

[16] I.F. Liyas, V. Markl, P. Haas, P. Brown, A. Aboulnaa, CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies, *in Proceedings of the 2004 ACM International Conference on Management of Data (SIGMOD),* France, 2004, pp. 647-658.

[17] ARFF, University of Waikato, Extensible attribute-Relation File Format, Available online from: http://weka.wikispaces.com/XML

[18] V. Pudi, P. Krishna, Data Mining, 1st ed., New Delhi, India: OXFORD University Press, 2009.

[19] F. Hakimpour, A. Geppert, Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach, ACM, Proc. of the Int. Conf. On Formal Ontologies in Information Systems FOIS, 2001, pp. 297-308.

[20] MSDN Hashing, Available Online: http://msdn.microsoft.com/en-us/library/system.string.gethashcode.aspx

[21] Y. Karasneh, H. Ibrahim, M. Othman, R. Yaakob, A model for matching and integrating heterogeneous relational biomedical databases schemas, in *Proc. of Int. Database Engineering & Applications Symposium*, Rende (CS), Italy, 2009, pp. 242 - 250.

[22] A. Fuxman, M. A. Hernandez, H. Ho, R. J. Miller, P. Papotti, L. Popa, Nested mappings: schema mapping reloaded, in *Proc. of the 32nd Int. Conf. on Very large data bases*, Seoul, Korea, 2006, pp. 67 - 78.

[23] R. Pottinger, P. A. Bernstein, Schema merging and mapping creation for relational sources, in *Proc. of the 11th Int. Conf. on Extending Database Technology,* Nantes, France, *2008*, pp. 73-84.

[24] The World Bank, Data Catalog, Available Online: http://data.worldbank.org/topic

[25] E. Rahm, P. A. Bernstein, A survey of approaches to automatic schema matching, in *the Very Large data bases Journal*, Vol 10 Issue 4, Dec 2001, pp. 334 - 350.

[26] A. Fuxman, E. Fazil, and R. Miller, ConQuer: Efficient Management of Inconsistent Databases, in *Proc. of the 2005 ACM SIGMOD Int. Conf. on Management of data,* Baltimore, Maryland, 2005, pp. 155 - 166.

[27] Z. Li, S. Li, Z. Peng, Ontology matching based on Probabilistic Description Logic, in *Proc. of the 7th Int. Conf. on Applied Computer & Applied Computational Science*, Hangzhou, China, April 2008.

[28] C. Yu, H. V. Jagadish, Querying complex structured databases, in *Proc. of the 33rd Int. Conf. on Very large data bases,* Vienna, Austria, 2007, pp. 1010-1021.

[29] T. Poulain, N. Cullot, K. Yetongnon, Ontology Mapping Specification in Description Logics for Cooperative Systems, in *Proc. of the 1st Int. Conf. on Signal-Image Technology and Internet-Based Systems,* 2005, pp. 240-246.

[30] The US Federal Government, Data Catalog, Available Online: http://www.data.gov/catalog

[31] H. Galhardas, D. Florescuand, An Extensible Framework for Data Cleaning, in *Proc. of the 16th Int. Conf. on Data Engineering,* California, USA, 2000, pp. 312.

[32] N. Choi, Y. Song, and H. Han, A Survey on Ontology Mapping, in *ACM SIGMOND Record*, Volume 35, Issue 3, 2006, pp. 34-41.

[33] The World Wildlife Fund, Data Catalog, Available Online: http://www.worldwildlife.org/science/data/item1872.html

[34] A. K. Elmagarmid, P. G. Ipeirotis, V S Verykios, Duplicate Record Detection: A Survey, in *IEEE Transactions on Knowledge and Data Engineering*, Volume 19, Issue 1, Jan 2007, pp. 1-16.

[35] P. Goiser, K. Christen, Quality and complexity measures for data linkage and deduplication, in *Quality Measures in Data Mining Book*, Volume 43, Springer, 2007.

[36] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, Adaptive Name Matching in Information Integration, in *IEEE Intelligent Systems Journal*, Volume 18 Issue 5, Sept 2003, pp. 16 - 23.

[37] G. Navarro, A Guided Tour to Approximate String Matching, in *ACM Computing Surveys Journal*, Volume 33, Issue 1, March 2001, pp. 31-88.

[38] J. Pan, C. Cheng, G. Lau, and H. K. Law, Utilizing Statistical Semantic Similarity Techniques for Ontology Mapping with Applications to AEC Standard Models, in *the Journal of Tsinghua Science & Technology*, Volume 13, pp. 217-222.

[39] N. Koudas, A. Marathe, D. Srivastava, Flexible string matching against large databases in practice, in *the Proceedings of the 13th Int. Conf. on Very large data bases,* Toronto, Canada, 2004, pp. 1078 - 1086.

[40] N. Brobergand, A. Farre, and J. Svenningsson, Regular Expression Patterns, in *the Int. Conf. on Functional Programming,* Utah, USA , 2004, pp. 67-68

[41] M. Ceci, A. Appice, C. Loglisci, D. Malerba, Complex objects ranking: a relational data mining approach, in *Proc. of the 2010 ACM Symposium on Applied Computing*, Switzerland, 2010, pp. 1071-1077.

[42] The Adventure Works Database, Data Catalog, Available Online: http://sqlserversamples.codeplex.com/

[43] A. Poggi, D. Lembo, D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati, Linking data to ontologies, in *the Journal on Data Semantics,* Heidelberg, 2008, pp. 133-173.

[44] M. Bilenko, and R. Mooney, Adaptive Duplicate Detection Using Learnable String Similarity Measures, in *the Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining,* Washington DC, USA, 2003, pp. 39 - 48.

[45] P. Costa, L. Mottola, A. L. Murphy, G. P. Picco, TeenyLIME: transiently shared tuple space middleware for wireless sensor networks, in *the Proc. of the Int. Workshop on Middleware for sensor networks,* Melbourne, Australia, 2006, pp. 43 - 48.

[46] M. G. Elfeky, V. S. Verykios, A. K. Elmagarmid, TAILOR: A Record Linkage Tool Box, in *the Proc. of the 18th Int. Conf. on Data Engineering*, California, USA, 2002, pp. 17.

[47] G. Noren, R. Orre, and A. Bate, A hit-miss model for duplicate detection in the WHO drug safety database, in *Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge discovery in data mining*, Chicago, USA, 2005, pp. 459 - 468.

[48] Statistics New Zealand, in *the Data Integration Manual,* Wellington, http://www.stats.govt.nz/~/media/Statistics/about-us/policies-protocols-guidelines/data-integration-further-technical-info/DataIntegrationManual.pdf, 2006.

[49] National Climatic Data Center, Data Catalog, Available Online: http://www.ncdc.noaa.gov/oa/ncdc.html

[50] Queensland Govt. Wildlife & Ecosystems, Data Catalog, Available Online: http://www.derm.qld.gov.au/wildlife-ecosystems/index.html

[51] W. E. Yancey, BigMatch: A Program for Extracting Probable Matches from a Large File, in *US. Census Bureau Research Report,* 2002

[52] A. Isaac, S. Wang, C. Zinn, H. Matthezing, L. van der Meij, S. Schlobach, Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain, in *IEEE Intelligent Systems Journal,* Volume 24 Issue 2, Mar 2009, pp. 76-86.

[53] H. Lee, R. T. Ng, K. Shim, Approximate substring selectivity estimation, in *the Proc. of the 12th Int. Conf. on Extending Database Technology,* St Petersburg, Russia, 2009, pp. 827-838.

[54] D. Calvanese, G. Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati, Linking Data to Ontologies: The Description Logic DL-Lite, in *the OWL: Experience and Direction Workshop,* Athens, Georgia, 2006,

[55] B. B Hariri, H. Sayyadi, H. Abolhassani and K. Sheykh Esmaili, Combining Ontology Alignment Metrics Using the Data Mining Techniques, in *Proc. of the 2006 Int. Workshop on Context and Ontologies,* Trento, Italy, 2006.

[56] Medical Science, Data Catalog, http://www.medicare.gov/download/downloaddb.asp

[57] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, 2006, Book ISBN: 3540331727 .

[58] S.C. Gupta, V.K.Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand & Sons, 2002

[59] C. Ding, X. He, K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization, in *Proc. of the 2004 ACM symposium on Applied Computing,* Nicosia, Cyprus, 2004, pp. 584 - 589.

[60] D. Balzarotti, P. Costa, G. P. Picco, The LighTS Tuple Space Frawework and its Customization for Context-Aware Applications, in *the Web Intelligence and Agent Systems Journal,* The Netherlands, Volume 5 Issue 2, Apr 2007, pp. 215-231.

[61] G. Cormode, A. McGregor, Approximation algorithms for clustering uncertain data, in *Proc. of the Int. Conf. on Principles of Database Systems,* Vancouver, Canada, 2008, pp. 191-200.

[62] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Qi Su, S E Whang, J Widom, Swoosh: a generic approach to entity resolution, in *the Very large data bases Journal*, Volume 18, Issue 1, Jan 2009, pp. 255 - 276.

[63] M.A. Hernandez and S.J. Stolfo, The Merge/Perge Problem for Large Databases, *In Proc of the ACM Int. Conf. on Management of Data (SIGMOD),* San Jose, California, May 1995.

[64] T. Churches, P. Christen, K. Lim and J. X. Zhu, Preparation of name and address data for record linkage using hidden Markov models, in *the BioMed Central Medical Informatics and Decision Makin Journal*, http://www.biomedcentral.com/1472-6947/2/9/, 2002.

[65] V. Borkar, K. Deshmukh, and S. Sarawagi, Automatic Segmentation of Text into Structured Records, in *Proc. of the 2001 ACM SIGMOD Int. Conf. on Management of Data,* Santa Barbara, California, 2001, pp. 175-186.

[66] Semantic Enrichment: The Key to Successful Knowledge Extraction, in *the Scope eKnowledge Center Literature*, Chennai, India, Available Online at: http://www.scopeknowledge.com/Semantic_Processingnew.pdf, Oct 2008.

[67] P. Christen, Febrl: a freely available record linkage system with a graphical user interface, in *Proc. of the Australasian Workshop on Health Data and Knowledge Management*, Canberra, Australia, 2008, pp. 17-25.

[68] A. Doan, J Madhavan, P Domingos, and A Halevy, Ontology Matching: A Machine Learning Approach, in *the Handbook on Ontologies in Information Systems,* Springer, 2003, pp. 397-416.

[69] P. Avesani, F. Giunchiglia, and M. Yatskevich, A Large Scale Taxonomy Mapping Evaluation, in *Proc. of the 4th Int. Semantic Web Conference*, Galway, Ireland, 2005, pp. 67-81.

[70] S. Ponzetto, and R. Navigli, Large-scale Taxonomy Mapping for Restructuring and Integrating Wikipedia, in *Proc. of the 21st Int. Joint Conf. on Artificial Intelligence,* Pasadena, USA, 2009, pp. 2083–2088.

[71] S. Muthaiyah, M. Barbulescu and L. Kerschberg, A hybrid similarity matching algorithm for mapping and rading ontologies via a multi-agent system, in *Proc. of the 12th WSEAS Int. Conf. on Computers*, Crete Island, Greece, 2008, pp. 653-661.

[72] Y. Zhai, and B. Liu, Web Data Extraction Based on Partial Tree Alignment, in *Proc. of the Int. World Wide Web Conference*, Chiba, Japan, 2005, pp. 76-85.

[73] Y. Ding, and D. Embly, Using Data-Extraction Ontologies to Foster Automating Semantic Annotation, in the *22nd Int. Conf. on Data Engineering Workshops*, Atlanta, USA, 2006, pp. 138.

[74] A. Laender, B. Ribeiro-Neto, A. Silva, and J. Teixeria, A Brief Survey of Web Data Extraction Tools, in *the ACM Int. Conf. on Management of Data* (*SIGMOD),* Volume 31, 2002, pp. 84-93.

[75] A. Shareha, M. Rejeswari and D. Ramchandram, Multimodal Integration (Image and Text) Using Ontology Alignment, in *the American Journal of Applied Sciences*, 2009, pp. 1217-1224.

[76] K. Munir, M. Odeh, R. McClatchey, Ontology Assisted Query Reformulation Using the Semantic and Assertion Capabilities of OWL-DL Ontologies, in *Proc. of the 2008 Int. Symposium on Database Engineering & Applications*, Coimbra, Portugal, 2008, pp. 81-90.

[77] A. Firat, S. Madnick, and B. Grosof, Knowledge Integration to Overcome Ontological Heterogeneity: Challenges from Financial Information Systems, in *Int. Conf. on Information Systems,* Barcelona, 2002, pp. 17.

[78] G. Járosa, Teleonics of health and healthcare: Focus on health promotion, *In World Futures: The Journal of Global Ed*, 54(3), Jun. 2010, pp. 259 – 284.

[79] M. Gollapalli, X. Li, I. Wood, and G. Governatori, Approximate Record Matching using Hash Grams, in *the IEEE International Conference on Data Mining Workshop*, 2011, Vancouver, Canada, pp. 504-511.

[80] D. McD Taylor, B. Bell, A. Holdgate, C. MacBean, T. Huynh, O. Thom, M. Augello, M., R. Millar, R. Day, A. Williams, P. Ritchie, and J. Pasco, Risk factors for sedation-related events during procedural sedation in the emergency department, Emerg. Med. Australasia. 23(4), Aug. 2011, pp. 466 – 473.

[81] F. Azam, Biologically Inspired Modular Neural Networks, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, PhD Thesis, 2000.

[82] R. Rojas, Neural Networks - A Systematic Introduction, 4th ed. New Yourk, U.S.A: Springer-Verlag, 2004.

[83] R. Bose, Knowledge management-enabled health care management systems: capabilities, infra-structure, and decision-support, Expert Syst. Appl. 24(1), Jan. 2003, pp. 59–71.

[84] Queensland Health. 2012, Protocol for Clinical Data Standardization, Document Number # QH-PTL-279-1:2012.

[85] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, D. Srivastava, Using q-grams in a DBMS for Approximate String Processing, *in IEEE Data Engineering Bulletin*, Vol 24, No. 7, Dec 2001.

[86] T. Turner, Developing evidence-based clinical practice guidelines in hospitals in Australia, Indonesia, Malaysia, the Philippines and Thailand: values, requirements and barriers. BMC Health Serv. Res. 9(1), Dec. 2009, pp. 235.

[87] E. E. Roughead, and S. J. Semple, Medication safety in acute care in Australia 2008, Aust. New Zealand Health Policy 6, Aug. 2009, pp. 18.

[88] A Fuxman, E Fazil, and R Miller, ConQuer: Efficent Management of Inconsistent Databases, *in ACM SIGMOD Int. Conf. on Management of Data* 2005, June 14-16, Baltimore, Maryland, p.p.155 - 166.

[89] Australian Government Department of Health and Ageing. Summary of the national E-Health Strategy: 2, National Vision for E-Health, Dec. 2009, pp. 1.

[90] M G. Elfeky, V S. Verykios, A K. Elmagarmid, TAILOR: A Record Linkage Toolbox, *in IEEE Int. Conf. on Data Engineering ICDE*, 2002, p.p.17-28.

[91] Australian Commission on Safety and Quality in Health Care. Safety and Quality Improvement Guide Standard 4: Medication Safety Oct. 2012, pp. 14.

[92] A. Burls, AGREE II—improving the quality of clinical care, The Lancet, 376(9747), Oct. 2010, pp. 1128 – 1129.

[93] J. Robertson, J. B. Moxey, D. A. Newby, M. B. Gillies, M. Williamson, and E. Pearson, Electronic information and clinical decision support for prescribing: state of play in Australian general practice, Fam Pract. 28(1), Feb. 2011, pp. 93-101.

[94] V. N. Stroetmann, D. Kalra, P. Lewalle, A. Rector, J. M. Rodrigues, K. A. Stroetmann, G. Surjan, B. Ustun, M. Virtanen, and P. E. Zanstra, Semantic Interoperability for Better Health and Safer Healthcare: Deployment and Research Roadmap for Europe, SemanticHEALTH project, a Specific Support Action funded by the European Union 6th R&D Framework Programme (FP6). DOI=http://ec.europa.eu/ : 10.2759/38514.

[95] W. Wenjun, D. Lei, D. Cunxiang, D. Shan, and Z. Xiankun, Emergency plan process ontology and its application, *In Proc. of the 2nd Int. Conf. on Advanced Computer Control (Shenyang, Liaoning* ICACC, Tianjin, China, 2010, pp. 513-516.

[96] M. Sotoodeh, Ontology-Based Semantic Interoperability in Emergency Management, Doctoral Thesis, University of British Columbia, 2007.

[97] Y. Peng, Application of Emergency Case Ontology Model in Earthquake, *in Proc. of Int. Conf. on Management and Service Science. MASS*, 2009, Tianjin, China, pp. 1 – 5.

[98] A. Bellenger, An information fusion semantic and service enablement platform: The FusionLab approach. Proceedings of the 14th International Conference on Information Fusion (Chicago, Illinois, USA, July 05 - 08, 2011), FUSION 2011. Val-de-Reuil, France, pp. 1 – 8.

[99] J. Hunter, P. Becker, A. Alabri, C. Van Ingen, E. Abal, Using Ontologies to Relate Resource Management Actions to Environmental Monitoring Data in South East Queensland, IJAEIS, 2(1) (2011), pp. 1-19.

[100] V. Mascardi, V. Cordi, P. Rosso, A Comparison of Upper Ontologies, Technical Report DISI-TR-06-2, The University of Genoa, The Technical University of Valencia, Genova, Italy, 2007.

[101] Suggested Upper Merged Ontology (SUMO), Available online at www.ontologyportal.org.

[102] Z. Xianmin, Z. Daozhi, F. Wen, and W. Wenjun, Research on SUMO-based Emergency Response Management Team Model, *in the Proc. of the Int. Conf. on Wireless Communications, Networking and Mobile Computing,* Shanghai, China, September 21 - 25, 2007, pp. 4606 - 4609.

[103] International Organization for Standardization, International Standard 18629, Available online at http://www.iso.org.

[104] M. Grüninger, T. Hahmann, A. Hashemi, D. Ong, and A. Özgövde, Modular first-order ontologies via repositories, Applied Ontology 7(2), April 2012, pp. 169 – 209.

[105] C. Lange, O. Kutz, T. Mossakowski, and M. Grüninger, The Distributed Ontology Language (DOL): Ontology Integration and Interoperability Applied to Mathematical Formalization, *In Proc. of the Conf. on Intelligent Computer Mathematics CICM*, Bremen, Germany, Springer, Heidelberg, 2012, pp. 463-467.

[106] M. Gollapalli, A Framework of Ontology Guided Data Linkage for Evidence based Knowledge Extraction and Information Sharing, *in the 29th IEEE International Conference on Data Engineering (ICDE) Workshop*, Brisbane, Australia, Apr 2013 [accepted on 20th Dec 2012]

[107] B. Bell, D. McD Taylor, A. Holdgate, C. MacBean, T. Huynh, O. Thom, M. Augello, R. Millar, R. Day, A. Williams, P. Ritchie, and J. Pasco, Procedural sedation practices in Australian Emergency Departments. Emerg, Med. Australasia. 23(4), May 2011, pp. 458 – 465.

[108] A. Holdgate, D. McD Taylor, B. Bell, C. MacBean, T. Huynh, O. Thom, M. Augello, R. Millar, R. Day, A. Williams, P. Ritchie, and J. Pasco, Factors associated with failure to successfully complete a procedure during emergency department sedation, Emerg. Med. Australasia. 23(4), Aug. 2011, pp. 474 – 478.