Enabling the Convergence of Traditional Electrical Regulatory Framework Documents and Smart Grid Functions Using Digitalisation

SETU
Ollscoil
Teicneolaíochta
an Oirdheiscirt
South East
Technological
University

David Ryan

School of Computing, Maths and Physics

South East Technological University

This dissertation is submitted for the degree of

Masters by Research

Supervisors: Bernard Butler, Brendan Jennings and Pádraig Lyons

Walton Institute for Information
and Communications Systems
Science

Submitted to South East
Technological University
September 2022

I would like to dedicate this thesis to my partner Mary and two boys, Adam and Conor, who supported and encouraged me with all their hearts.

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Masters by Research, is entirely my own work and has not been taken from the work of others save to the extent that such work has been cited and acknowledged within the text of my work.

David Ryan
Submitted to South East Technological University September 2022

Acknowledgements

During the course of this journey there have been some people who have had a huge impact on its direction, destination and on the quality of the road. From the outset my supervisor Dr Bernard Butler has been involved in this journey and his impact on the quality, direction and supervision cannot be over stated. His guidance both kept me both grounded in the research and pushed me to expand my horizons in equal measure. His advice and comments when writing this thesis were invaluable and I owe him a debt of gratitude. I also want to thank Miguel Ponce de Leon who, as my manager at the time of carrying out this research, afforded me the time to carry out this research and who also acted as a sound board on more than one occasion to hear my ideas and provide some feedback. I also express my gratitude to Dr Pádraig Lyons from ESB Networks who provided some industry perspective at the early stages of the project and also provided valuable review comments when completing this thesis. I would also like to thank my colleagues in the TSSG in general who potentially and unwittingly helped me in some shape or form, whether that be sending me a link to some relevant research, gave their opinion on something I was trying or reviewed something for me. I would also like to acknowledge the contribution of the examiners, Prof Andrew Keane and Dr Robert Brennan, who scrutinised the research with an expert eye from both the electrical and text mining fields respectively, and through their comments made a positive impact on this dissertation. I want to also extend a word of thanks to my parents, Michael and Mary Ryan, who from an early age instilled in me work ethic and an appetite to learn, qualities that were essential in equal measure throughout this journey. All of those mentioned above have helped shape this research and helped me through this journey and I will be eternally grateful.

Abstract

With concerted efforts by legislators to reduce CO2 emissions, the ways in which we produce, regulate, and consume power is going through wholesale changes. These changes will impact the utilities as they are central components in this shift and will be required to balance ensuring grid stability against allowing distributed energy sources, which are less stable than traditional generation, to participate in the supply of "green power" to the grid. Academia and industry are combining resources to tackle this challenge with advances in power systems electronics methods, applying advanced ICT technologies and testing new regulatory and market frameworks and models. This is evident in the Electrical Industry Regulatory Policies or Network Codes, which is a set of agreed rules between the actors that participate in the supply, distribution, transmission, and regulation of electrical power. In Ireland, like many other countries, these codes are contained in static documents and consist of quantifiable constraints, interspersed with text. The only way to verify, update or interpret the codes is by humans reading and editing them directly. The work in this thesis, using cloud-based text mining and graph database management, aims to extract the network codes from the documents and derive a richer representation of the network codes, while maintaining their providence and structure, consistent with their representation in the document. This work derived from two overarching scenarios which encompass the research questions, motivate the research methodology and help explore the potential impact that the proposed enriched representation might have on the business processes within the electricity supply industry. These scenarios are derived from real challenges faced in the industry and were selected because of their practical interest for energy stakeholders but we also consider the technical and business potential of our proposed extraction and representation processes.

Contents

List of Figures

List of Tables

List of Acronyms

CIM Common Information Model. 17, 20, 21, 68

DER Distributed Energy Resource. 4, 6, 67
DSO Distribution Supply Operator. 1, 5, 8
DSU Demand-Side Unit. 7

GIS Geographical Information System. 17, 20, 21, 68

ICT Information and Communication Technology. 1, 2, 11, 12, 17
IoT Internet of Things. 1, 3

JSON JavaScript Object Notation. 20, 73

LV Low Voltage. 3

MQTT Message Queuing Telemetry Transport. 17
MV Medium Voltage. 3

NLP Natural Language Processing. 10, 45, 51

OCR Optical Character Recognition. 14–16

PDF Portable Document Format. 15

RES Renewable Energy Source. 1, 2
REST Representational State Transfer. 17, 73

SCADA Supervisory Control and Data Acquisition. 16, 21
SEM Single Electricity Market Operator. 5

TF-IDF Term Frequency-Inverse Document Frequency. 51
TOC Table of Contents. 76

TSO Transmission Supply Operator. 1, 5, 6, 8

VPP Virtual Power Plant. 67

XML Extensible Markup Language. 15–17

Chapter 1

Introduction

This chapter outlines the work in this thesis by describing the challenge being faced by current regulatory frameworks in the smart grid from a semantic alignment perspective. To gain an understanding and assess the scope of the challenge, it is essential that the relevant components are understood, so the background of the key contributors of the challenge are presented. Next an explanation is given describing the key contributions of this thesis that can enable the development of solutions that can meet this challenge. The remainder of this chapter outlines the organisation of this thesis.

1.1 Challenge

The electrical supply industry is undergoing many changes, from how the network is monitored and managed, to how energy is consumed as a commodity and how the relationships between new and old stakeholders in the market are evolving. The traditional structure of the electrical supply industry was based on mature, centrally controlled providers fed by stable predictable generation sources and governed by mature long standing relationships between stakeholders (generators, Transmission Supply Operator (TSO)s, Distribution Supply Operator (DSO)s and retailers).

The change is driven by the need to reduce CO_2 emissions by increasing Renewable Energy Source (RES) usage which in turn is causing a move away from analog-based grid control systems towards digitalisation. The resulting Smart Grid combines Internet of Things (IoT) sensors, advanced Information and Communication Technology (ICT) infrastructure and software and improved power electronics to move from a centrally controlled to a more distributed system. This convergence of power systems electronics and ICT, can provide stable and resilient electrical networks, but the centralised processes now used to manage and regulate power and data flows through the electrical network may no longer be optimal. According to Kumar et al. (2012), the existing regulatory frameworks have served the electrical industry since the inception of large scale electrification and the subsequent deregulation of electrical power generation, but the frameworks need to be reconsidered to match the needs of the Smart Grid, where the number of administrative domains is much greater than before.

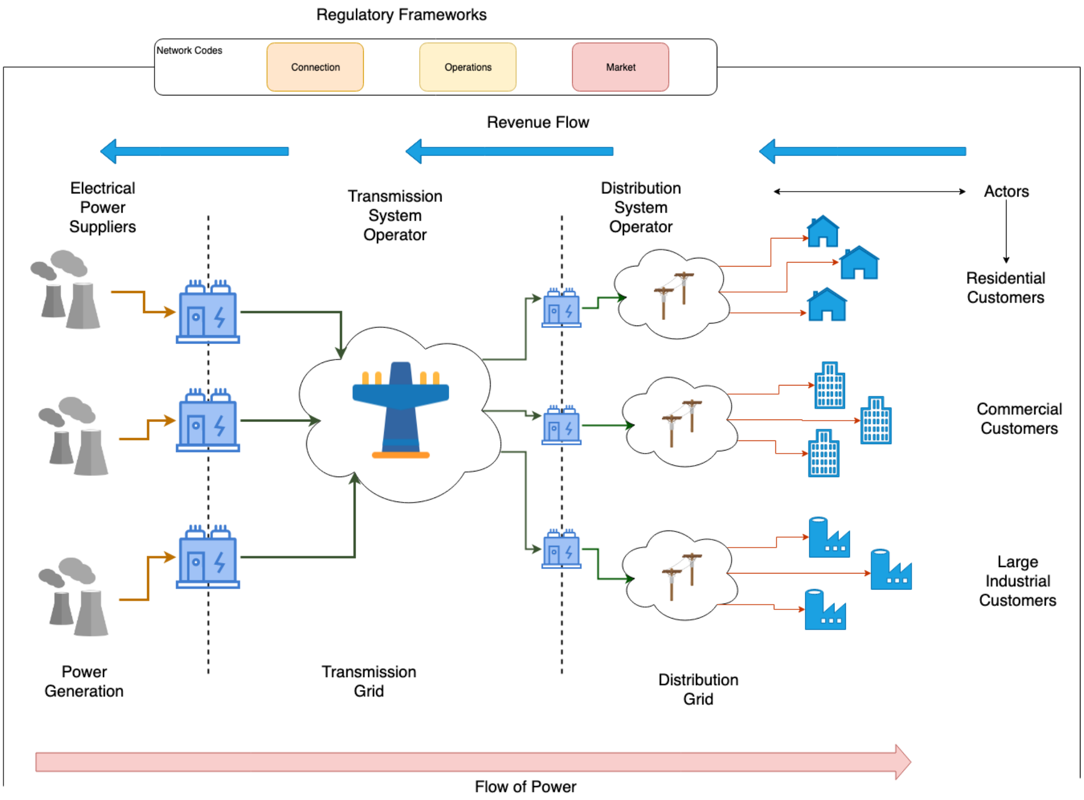


Fig. 1.1 The traditional grid depicting traditional actors and unidirectional power and revenue flows

Figure 1.1 is a representation of the traditional grid that serves to highlight the unidirectional nature of the power and revenue flows, the clearly defined role of each actor, the simplicity of the high level topology of the electrical system and the heterogeneous synchronous nature of the power supply. Contrast that with the modern or emerging grid representation in figure 1.2 where there is a more diverse set of actors that are looking maximise the revenue generated by their RES assets, a more complex grid electrical system topology with the introduction of power generation at the Low Voltage (LV) and Medium Voltage (MV) levels and diversity in terms of the power supply. These factors are changing the nature of the power and revenue flows from unidirectional to bidirectional, resulting in the system moving further towards or in some cases exceeding its operational limits and also leading to the grid becoming more complex with the deployment of IoT devices, smart inverters and other components that enable the digitalisation of the electrical power system. While these figures are simplistic and general in terms of the market layout, with more actors, like Market Operators and Energy Retailers, and more complex cash flows between actors, they really serve to highlight the general flow of power and data.

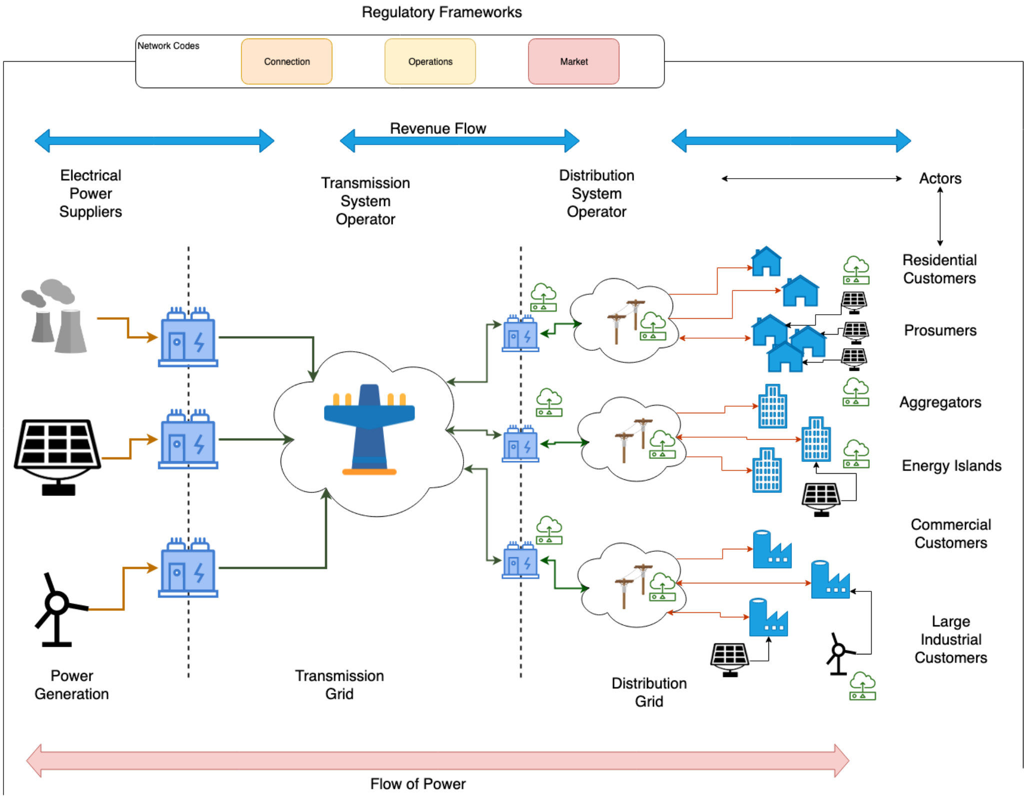


Fig. 1.2 The modern grid depicting traditional and new actors and bidirectional power and revenue flows

At the top of both diagrams are the regulatory frameworks and these contain the Network Codes, which include Distribution Codes, Grid Codes and the EU Network Codes, and while the grid is going through the changes outlined, the Network Codes have not changed in terms of their level of digitalisation. The Network Codes cover a broad range of topics such as connection, regulation and market guidelines and all versions have broadly the same assertions and requirements but they are not specified in the same way. Both the notation and the detailed semantics can differ, so human interpretation is often needed to determine whether there is functional semantic alignment. Furthermore, they have dependencies expressed as references to other regulatory documents, including those published by the same body (and hence more likely to share similar terminology) and externally (at national and EU level, say). While this research concerns documents written in the English language the challenge of multilinguality is also present. These problems (of proving semantic alignment and of managing links to a variety of other reference documents) present difficulties when updating existing codes or creating new codes. Such difficulties show up in the wide variety of network code and dependent document formats, and even the duplication of nominally equivalent parameters and codes across all the relevant stakeholders. The proliferation of Distributed Energy Resource (DER), immature stakeholder relationships and piecemeal attempts at digitalisation form barriers to progress towards having dynamic, updatable and semantically aligned Network Codes, especially when there is a diversity of supply as an increased number of smaller suppliers enter the market, collectively offering energy that was generated with lower carbon emissions.

1.2 Background

The Challenge identified the changes that the electrical industry is undergoing and specifically the gap in the levels of digitalisation between the smart grid and the network codes. This section explores further the background of the electrical industry in the context of digitalisation and structure and also details the background of text mining as it forms part of the solution investigated in this research to bridge the digitalisation gap.

1.2.1 The Electrical Industry

1.2.1.1 Traditional and Emerging Actors in the Electrical Industry

According to Gaffney et al. (2017), shortly after the foundation of the Irish state, the Electricity Regulation Act was created in 1927. This saw the Electricity Supply Board being established to manage the generation, transmission and distribution of power to the nation and with it the first organised electricity market in Ireland. Until the establishment of the state in 1922, Ireland was a political constituent of the United Kingdom and the shift towards electrification mirrored the electrification there. According to the ESB Archives[1], 400,000

[1] https://esbarchives.ie/key-rural-facts/

rural homes in Ireland in 1946 were without electricity, with the majority of the electrification delivered to urban homes. Between then and 1964 saw the introduction of the Rural Electrification Scheme erect 1 million poles carrying 50,000 miles of line delivering power to 300,000 homes. The scheme continued in phases and by 1975 99% of the country had electricity across 792 rural areas that were divided into 10 districts. Since Ireland joined the European Union (previously the European Economic Community) in 1972, the sector here has been influenced by directives in the form of energy packages and these directives helped shape the modern electricity industry in Ireland today.

The modern electrical industry comprises four key functional areas: Generation, the Transmission System, the Distribution System and the Single Energy Market. The management of each of these functional areas differs from country to country but the actors include Distribution Supply Operator (DSO), Transmission Supply Operator (TSO), Single Electricity Market Operator (SEM) and Generation companies. These actors have clearly defined roles and responsibilities within the industry, with the societal goal of delivering a well-regulated, stable, secure, affordable and safe supply of power to the consumer.

As defined by E-DSO (E.DSO, 2020), the European Distribution System Operators group, the

> Distribution system operators (DSOs) are the operating managers (and sometimes owners) of energy distribution networks, operating at low, medium and, in some member states, high voltage levels (LV, MV, HV). By contrast, transmission grids transport large quantities of high (and extreme high) voltage (HV, EHV) electricity across vast distances, often from large power plants to the outskirts of large cities or industrial zones, where it is transformed into lower voltages for DSOs to distribute to end-users through the distribution network. Overhead and underground cables leading to homes and businesses are operated by DSOs.

In Ireland, ESB Networks Ltd is the sole DSO but in other countries there can be many DSOs and in some cases the DSO might not be a commercial entity but might, as is the case in Nordic regions, be a local authority.

As described by ENTSO-E (ENTSO-E, 2020), the European Network of Transmission System Operators for Electricity,

> TSOs are entities operating independently from the other electricity market players and are responsible for the bulk transmission of electric power on the

> main high voltage electric networks. TSOs provide grid access to the electricity market players (i.e. generating companies, traders, suppliers, distributors and directly connected customers) according to non-discriminatory and transparent rules. In order to ensure the security of supply, they also guarantee the safe operation and maintenance of the system. In many countries, TSOs are in charge of the development of the grid infrastructure too.

Similarly to ESB Networks, Eirgrid plc [2] is the sole TSO in the Republic of Ireland with responsibility for the entire transmission system for the Republic of Ireland and also, in conjunction with some independent TSOs like OFREG[3], it has a part to play in the management of the transmission system in Northern Ireland.

The role of the Generation companies in the Electrical Industry is to provide power of an acceptable quality and quantity to the transmission system. These Actors were traditionally large companies, sometimes the DSO or a subsidiary of a DSO, that provide large generation plants that were traditionally run on fossil fuels, nuclear or hydroelectric power. This landscape is changing, however, with the emergence of micro generation units and DERs. New entrants come into this market each year, increasing the supply of renewable energy. However, it is still the case that the main energy suppliers in Ireland are ESB Generation [4] and SSE Generation [5].

According to the Irish Department of Communications, Climate Action and the Environment, Department of Communications and the Environment (2020)

> The Single Electricity Market (SEM) is the wholesale electricity market for the island of Ireland. It is regulated jointly by the Commission for Regulation of Utilities (CRU) and its counterpart in Belfast, the Utility Regulator. The SEM combined what were two separate jurisdictional electricity markets. The SEM became one of the first of its kind when it went live on 1st November 2007. The goal of the SEM to provide for the least cost source of electricity generation to meet customer demand at any one time across the island, while also maximizing long-term sustainability and reliability.

As can be seen in 1.2 are other Actors emerging in the industry such as Aggregators, which are downstream of independently owned micro-generation units, and aggregate power that is

[2] http://www.eirgridgroup.com/

[3] https://www.ofreg.ky/

[4] https://www.esb.ie/tns/education-hub/electricity-generation

[5] https://www.sserenewables.com/

then provided to the grid for general usage. An example of this would be system called a Demand-Side Unit (DSU), as described by (Eirgrid, 2020) as consisting “of one or more Individual Demand Sites that can be dispatched by the Transmission System Operator (TSO) as if it was a generator”, where aggregators provide a DSU that monitors the power usage for premises with self-generation and at times of low energy consumption the unit allows the power to be exposed to the grid. The role of the aggregator here would be to aggregate and manage the virtual power exported and trade this on the market.

Energy Service Companies (ESCOs), as defined by the ESRI (2000), are commercial or non-profit businesses providing a broad range of supporting services to the main actors in the Irish energy supply system, including the design and implementation of energy saving projects, retrofitting, energy conservation, energy infrastructure outsourcing, power generation and energy supply, and risk management.

To ensure the fair and balanced operation of the power system and market there is an independent regulator. This is a public body that has jurisdiction over all entities within the energy supply industry and defines, validates, ratifies and enforces policies and regularity frameworks that govern the electricity area as a commodity. In the Republic of Ireland this is carried out by the Commission for Regulation of Utilities[6], a government appointed independent agency responsible for the responsible for the regulation of energy and water with a range of economic, customer care and safety functions.

1.2.2 Electrical Industry Regulatory Policies

The supply of electrical power is a heavily regulated sector and these regulations are contained in a suite of documents that cover such regulatory areas as:

- Connection
  - Requirements For Generators
  - Demand Connection Code
  - High Voltage Direct Current Connections
- Operations
  - System Operations
  - Emergency and Restoration

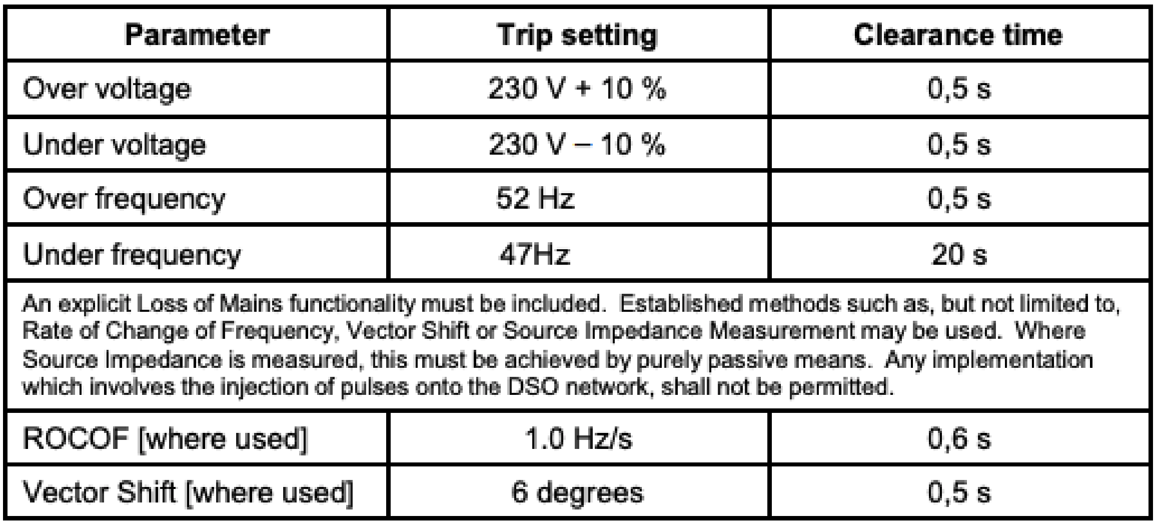
[6] http://www.cru.ie

- Market
    - Capacity Allocation & Congestion Management
    - Forward Capacity Allocation
    - Electricity Balancing

While these areas appear to map directly on to the functional areas in the electricity supply industry, this is not the case in practice and from a TSO and DSO perspective in Ireland, these regulations are copied across both Actors and comprise many different sub-policies and standards. The DSO version, named the Distribution Codes (ESB, 2016), comprise not only Distribution Codes but also the TSO's Grid Codes (EirGrid Grid Code, 2015) and Network Codes. In comparison to the Distribution codes, which contain parameters that control components in the high-level regulatory areas above and also act as a source of standards that govern hardware capabilities and electrical engineering practices (Lyons et al., 2018), the TSO versions are more high level. The DSO versions share some of these high level codes but also contain more specific codes that centre on the customer interface and specific distribution system asset management. This mirrors the level of complexity when we compare the TSO and DSO in figure 1.1 and note the number of devices and connections.

Upon visually inspecting the documents each code is structured with a code identifier like, CC.1.1, and a block of text and therefore it makes sense to break down the composition of these policy documents by parsing them into keys (the code identifier), values and supplementary text. The values are numeric parameters that control the distribution and transmission of power and the supplementary text provides context to the key-values so that they can be interpreted by humans. This informal text is designed to aid human readability. The need to align text and settings, the shared and overlapping nature of the documents, their interdependence and their dependence on external regulations makes alignment difficult. Given the relatively large number of settings embedded in a myriad of documents, this needs to be done in a scalable way. At present the management and updating of these requires manual intervention by experts in the TSO, DSO and regulator organisations to ensure seamless alignment across all versions, sub-documents and standards. Such experts draw upon their domain experience to overcome the lack of machine-readable/formal semantics in these documents and an inability, without expert knowledge, to gauge the impact that changing a value, say Reactive Power limits, in one place has across all interlinked codes, documents and standards (Lyons et al., 2018).

The following use case motivates our study of semantic alignment, as it explores the linkage between the Grid Codes, Distribution Codes and the EN50438 standard mandated by the SEAI (SEAI, 2013). EN50438 is a European standard that defines how micro-generators should be connected to the distribution network. There are some deviations to the EN50438 at a national level and these deviations are broadly governed by policy documents, such as a subset of the distribution codes and or a derivative of the Network Codes. The EN50438 standard forms, at a regulatory level, a certificate of compliance that all inverters for micro-generation must have before being considered for connection to the distribution network. Table 1 in (Networks, 2009), shown in figure 1.3, contains the Irish-specific variations to EN50438 and the details here consist of constraints regarding voltage and frequency. The constraints from a DSO perspective are referenced in Distribution Code DCC10.5.1 (ESB, 2016), which states that the highest voltage allowed at 230V level is 253V and this translated to the "Over voltage" parameter below is 230 V + 10%. Similarly, the Trip Setting for the ROCOF parameter, which is the rate of change of frequency, is laid out in section CC.7.3.1.1, which governs the Connection Condition for Generators, and references section 1.2 in a decision paper by the Commission for Energy Regulation (CER) (for Energy Regulation, 2014), CER14081-ROCOF, by the regulator that governs the Rate of Change of Frequency. It must be noted however that, at present, in network operations there is only ever one version of the code used by any one actor at any one time. The challenge being highlighted is the maintainability of multiple versions of the codes in multiple documents that must change simultaneously in a managed way with in multiple organisations for them to operate effectively. This use case highlights the problem: there are many different types of policy documents, value constraints can be expressed in different ways and so semantic matching needs to take account of all of them when deriving a federated, semantically aligned view of the relevant documents. Indeed, interpreting the constraints often depends on interpreting the supplementary text, so this also complicates matters. Lastly, it is also difficult to determine the ‘single version of the truth‘ if and when there is a circular dependence between value constraints in different documents.



Parameter	Trip setting	Clearance time
Over voltage	230 V + 10 %	0,5 s
Under voltage	230 V – 10 %	0,5 s
Over frequency	52 Hz	0,5 s
Under frequency	47Hz	20 s
An explicit Loss of Mains functionality must be included. Established methods such as, but not limited to, Rate of Change of Frequency, Vector Shift or Source Impedance Measurement may be used. Where Source Impedance is measured, this must be achieved by purely passive means. Any implementation which involves the injection of pulses onto the DSO network, shall not be permitted.		
ROCOF [where used]	1.0 Hz/s	0,6 s
Vector Shift [where used]	6 degrees	0,5 s

Fig. 1.3 EN50438 ESB Networks Representation

1.2.3 Text Mining

Text mining is a growing field and plays a major but often hidden role in many aspects of life. Its growth has been driven by digitalisation, is enabled by hardware and software advances and is motivated by the need to gain competitive advantage from the massive increase in unstructured textual data from social network, e-commerce and web based platforms. Most of this text is, unlike structured data which is managed using traditional database systems, unstructured and lacks the metadata to extract and relate key concepts held within. In humans we have an innate ability to apply and distinguish patterns in text using linguistic clues, something that computers struggle to handle because they lack contextual understanding, and can be distracted by differences in dialect, spelling variations and slang. Miner et al. (2012) outline three ways to overcome these challenges, i) text summarisation and classification, ii) information science and iii) natural language processing. These three techniques form the building blocks of most modern day text mining applications. Work in these areas are not particularly new and it is written that early examples of summarisation and classification date back to a library catalogue developed by Thomas Hyde for the Bodelian Library in the University of Oxford in 1674. Furthermore the foundations of information science were laid in the 1940s by Claude Shannon's seminal paper on information theory (Shannon, 1948) and by research by Skinner (1957) and Luhn (1958) in the 1950s, which was influenced by the philosophy of Aristotle and Plato, and work in the 1980s by Chomsky (1959) forming the basis of modern Natural Language Processing (NLP). The early work has formed a strong theoretical basis for text mining but true strides have been made in text mining, in practical terms, due to advances in hardware and software due to the development of High Performance Computing and Artificial Intelligence. Humans, although their understanding of language is far more advanced than that of a computer, which only has the capability to run programs to process and analyse text, they lack the capability process large swathes of data quickly, this is evident in the context of the Network Codes where a team of experts, while very familiar with the regulatory frameworks and the contents of the documents, would be required to sift through them manually to assess and perform the changes when required to do so.

1.3 Contribution of this Thesis

Chapter 1.1 (Challenge) describes how the increasing levels of change in grid systems and the growth of digitalisation are challenging the existing policies and standards in terms of the need to cater for new relationships, advances in power system electronics, the convergence of the fields of ICT and electrical engineering and the added complexity being introduced into the regulatory frameworks by the deployment of DER and micro-generation. This dissertation will explore the existing regulatory frameworks by defining the scope of the challenge they face in keeping pace with the Smart Grid and presenting mature ICT techniques to afford the existing regulatory frameworks the flexibility to support increasing digitalisation. From the perspective of the electrical industry expert the contribution of this thesis is to use ICT to enable flexible management of the process of specifying and agreeing electrical policies and standards to allow the industry expert better:

(a) align versions between organisations;

(b) assess the impact of changes in a single grid code document;

(c) assess the impact of changes across grid code documents spanning other relevant organisations;

and by doing so remove some of the time-consuming and labour-intensive manual processes currently in place when altering, validating and comparing policies and standards.

For ICT specialists, this dissertation provides a better understanding of the role text mining and concept modelling can play in streamlining processes in the smart grid, by reducing the scope for inconsistency between the different network codes and doing so in a way that scales well as more micro-suppliers join the grid. Furthermore the aim is to explore and define a process that will help match codes across documents that share the same domain without the presence of a corpus.

Summarising, the main contribution of this thesis is to use the convergence of ICT and power systems to benefit two main entities, the ICT specialist and the power industry expert. It will enable ICT specialists to better understand the network codes and their formats and based on the state of the art in text mining and concept modelling research to develop techniques to:

(a) Ingest the grid code documents;

(b) Convert the network codes from text based formats to a computer readable format;

(c) Normalise the grid code text;

(d) Extract the network codes as concepts;

(e) Assess the impact that a change in one code or related annex item would have on the entire set of network codes.

It will not further the state of the art in text mining or semantic alignment but will explore their application in the electrical industry, opening the potential for finding novel use cases, scenarios and system integration opportunities.

1.4 Organisation of this Thesis

This dissertation is structured as follows.

In this chapter we discussed the Background, the Challenges and how the Contribution of this Thesis aims to meet some of these challenges. Chapter 2 (Literature Review) presents the state of the art relating to The Digitalisation of Traditional Documents, Digital Knowledge Representation in the Electrical Supply Industry and Graph Similarity, Matching and Recommender Algorithms in Digital Information are currently achieved. To align with the challenge posed and the contribution of this thesis, Chapter 3 (Research Context and Scope) is defined by a Hypothesis and a set of Research Questions that frame the Research Methodology and their results. A set of Tools and Technologies (Appendix Item A) is also presented that explains the software tools and the software architecture used to run the experiments. The Conclusions in Chapter 5 present a Summary of the work carried out and detail the potential of this approach in the form of Future Work.

Chapter 2

Literature Review

In order to answer the research questions posed in Section 3.2 it is important to explore the state of the art in how traditional documents are handled in a digital age, the role semantic alignment has to play in smart contracts, how graph similarity is determined in digital information and also how digital information is represented in the smart grid. The findings of this literature review will help inform the experiments carried out in Chapter 4 and the tools and technologies chosen in Appendix Item A.

2.1 The Digitalisation of Traditional Documents

To fully explore the state of the art in the digitalisation of traditional documents we first must detail the fundamental differences between the digital and and traditional representation of a document. In a paper by Buckland (1998) it is stated that “A paper document is distinguished, in part, by the fact that it is on paper. But that aspect, the technological medium, is less helpful with digital documents.” and he goes on to say that a digital document is harder to define due to the various forms of media and broad range of applications that can generate textual information. In an article by Furuta (1995), they look at the differences between the traditional and digital and note three key differentiation factors,

- *permanence, if it is not on paper is it real,*
- *importance, is a paper diploma more important than a digital one*
- *size, the scope of a digital document has the greater scope to grow over a traditional one.*

Along with these abstract differences there are the obvious physical differences with one traditionally being a virtual representation and the other being stored in a computer readable format. These differences are fundamental to how we view and treat both types and even more important in how we convert the traditional to the digital. In a conference paper by Norrie and Signer (2003) they explore traditional documents as a relevant form of client device and they present an architecture for interactive paper called Paper++[1] that enables the user interact with the paper digitally by having links embedded into the paper which can be read and used by a digital device and served up to the user as hypertext. This approach might be seen as a step along the path to full digitalisation of documents. Quite often a traditional document may not be of the structured kind in terms of clearly delineated textual boundaries, and in a paper by Liang et al. (2005) a study was performed that surveyed the hardware technologies like Optical Character Recognition-based image capture and weighed up the pros and cons of using camera-based acquisition over using scanner technology to capture the document. They state that the composition of the document is a key factor in the success of any text extraction method but stress that at the acquisition phase that the text is still only an image. They present a set of steps for processing the captured image with the aim to extract text, text detection, localisation, extraction, geometrical normalization, enhancement/binarisation, and recognition. These steps are reliant on image processing and potentially still a very relevant process where documents concerned are, for example, “decaying, unique sources (no copies exist elsewhere) of personal information that constitute the only record and proof of existence for many thousands of people during the dark years of Nazi occupation in Europe” as outlined as a challenge in a paper by Antonacopoulos and Karatzas (2004) or as a method to extract text from handwritten historical documents. To be more specific in terms of the network codes it is worth noting that the network code documents are written using a text editor like MS Word and are presented and shared in PDF format, which are indeed digital representations, and therefore image processing technologies may not be necessary. However they have been written with Human Readability in mind over the use of their content for anything other than human referral and verification.

While the above describes techniques and research into the digitalisation of pre-existing documents and given that the subject matter for the experiments, the regulatory documents, Portable Document Format (PDF) documents generated from Microsoft Word files, it is worth exploring both these formats and how they can be exported and converted to other file

[1] https://beatsigner.com/paperpp.html

formats. In 1985, Microsoft introduced their first word processor called Multi-Tool Word, later to become Microsoft Word (MS Word) (Allan, 2001) that remains today the most widely-used word processing application (Chakravarty et al., 2006). MS Word allows the creation of editable and exportable documents in which multimedia items can be embedded. It allows for and provides formatting capabilities and is underpinned by utilising Extensible Markup Language (XML) (Tyson, 2007) which allows for the embedding of meta data like authorship and creation data and MS Word Version. As stated the subject matter for the experiments are PDF documents, exported from MS Word, and according to Adobe, the creators of PDF, "PDF files are built from a sequence of numbered objects similar to those used in the PostScript language" (Bienz et al., 1993). While it is possible to annotate PDF documents by adding text boxes, signatures and comments it is not possible to edit the content within. Core to enabling the use of the contents of the PDF documents is the ability to extract the textual content from them and there are several systems that carry out such tasks in terms of conversions to XML (Déjean and Meunier, 2006) and to HTML (Jiang and Yang, 2009) using techniques such as text detection and OCR (Dori et al., 1997). While academia has been active in these fields the open source software community has been most active creating software modules like the Javascript based[2] and Python's[3] based versions of PDF2Text and online tools such as PDF to TEXT[4] and PDF2GO[5] that effectively extract the text and metadata.

Since the regulatory documents typically take the form of PDF documents generated from Microsoft Word files, it is worth exploring both these formats and how they can be exported and converted to other file formats. In 1985, Microsoft introduced their first word processor called Multi-Tool Word, later to become Microsoft Word (MS Word) (Allan, 2001) and remains today the most used word processing application (Chakravarty et al., 2006). MS Word allows the creation of editable and exportable documents in which multimedia items can be embedded. It allows for and provides formatting capabilities and is underpinned by utilising Extensible Markup Language (XML) (Tyson, 2007) which allows for the embedding of meta data like authorship and creation data and MS Word Version. As stated the subject matter for the experiments are PDF documents, exported from MS Word, and according to Adobe, the creators of PDF, 'PDF files are built from a sequence of numbered objects similar to those used in the PostScript language' (Bienz et al., 1993). While it is

[2] https://github.com/robgraeber/pdf2text

[3] https://github.com/jalan/pdftotext

[4] https://pdftotext.com/

[5] https://www.pdf2go.com/pdf-to-text

possible to annotate PDF documents with text boxes, signatures and comments it is not possible to edit the content within. Core to enabling the use of the contents of the PDF documents is the ability to extract the textual content from them and there are several systems that carry out such tasks by converting to XML (Déjean and Meunier, 2006) and to HTML (Jiang and Yang, 2009) using techniques such as text detection and OCR (Dori et al., 1997). While academia has been active in these fields the open source software community has been most active creating software modules like the Javascript based[6] and Python's[7] based versions of PDF2Text and online tools such as PDF to TEXT[8] and PDF2GO[9] that effectively extract the text and metadata.

2.2 Digital Knowledge Representation in the Electrical Supply Industry

To create a solution for the representation of the digitised network codes it is important to investigate current methods of digital knowledge representation in the electrical supply industry. As interoperability is a key concern in the Smart Grid at present (Kim et al., 2017) it is essential that any novel representations of grid related functions or entities will not provide another interoperability concern. To achieve this it is necessary to explore the information systems that constitute the Smart Grid, the digital data they generate, how that data is transferred and represented and the standards and regulations that govern them. These information systems perform specific functions in the management of the power system from the generation, transmission and distribution of power to the control and monitoring of demand and usage from the consumer side. In modern grid systems a Supervisory Control and Data Acquisition (SCADA) (Boyer, 2009) system collects data from utility field devices, then uses it to manage the electrical grid infrastructure. It operates by analysing the data gathered from SCADA Remote Terminal Unit (RTU) (Osburn III, 2003) and based on that analysis can trigger preprogrammed controls or operator specific alarms. SCADA data is historically persisted and in most cases SQL databases are used, however, there is no defined schema for the storage of SCADA data and such schemas are driven by the business case. Outage Management Systems (OMS) (Jäger et al., 2016) need to visualise and interpret outages that can happen, in order to take corrective actions, minimize the effect, diagnose the

[6] https://github.com/robgraeber/pdf2text
[7] https://github.com/jalan/pdftotext
[8] https://pdftotext.com/
[9] https://www.pdf2go.com/pdf-to-text

causes and improve the system's availability and reliability. Geographical Information System (GIS) (Ghosh et al., 2013) is an important data representation for utilities as it provides them with a visualization of maps and points of interests and aids in the management and presentation of spatial data. The aforementioned are primarily grid management tools that typically enable the monitoring and control of the grid at High and Medium Voltage levels.

To create full visibility of the network, systems like Advanced Metering Infrastructure (AMI) (Rashed Mohassel et al., 2014), which allows the utility measure consumption and production of power at the lower levels of the network, and and systems like Demand Response Management System (DRMS) (Koch, 2014) which provide the utilities the ability to create automated, integrated, and flexible platforms to manage demand response solutions in an efficient and smart manner. While these information systems can be standalone there is an ongoing impetus within research and industrial circles to provide common standards and data integration mechanisms that ensure the consistent sharing of between systems in electrical networks. Some of these standards and approaches are founded in the communications standards like IEEE 1815 (DNP3) [10] and IEEE 2030.5 (Sep2) [11] which allow these systems use existing communications infrastructures while others like IEC 61850 (Samitier, 2017), which is a Meter Data Management System communications protocol and IEC 61970/61968 a Common Information Model (CIM) (Specht and Rohjans, 2013) based on Unified Modelling Language (UML) [12] that facilitates the normalisation and standardisation of the data between smart grid systems. In modern grid systems bi-directional data transfer between components, both cloud and edge device, is becoming a common feature. This bi-directional communication is underpinned by ICT communications protocols like Representational State Transfer (REST) and Message Queuing Telemetry Transport (MQTT) among others. Core to these technologies are standardised data communications formats like XML (Bray et al., 2008), which is flexible with the added option of combining it with an XML schema to enable stricter operations (Kanayama et al., 2014), and JSON (Crockford, 2006) which is lighter and less strict (Erfianto et al., 2007). Both are used in smart grid operations with the use of JSON becoming more prevalent due to the emergence of edge processing and the proliferation of memory-constrained devices at the edge of smart grid networks. In recent times there has been some research that centres on viewing the electrical network database as a graph database and viewing the topology of the

[10] IEEE Standard for Electric Power Systems Communications – Distributed Network Protocol (DNP3)

[11] IEEE Standard for Smart Energy Profile Application Protocol

[12] https://www.uml.org/

power system as a set of nodes and edges. For example, Anwar and Mahmood (2016), have viewed the bus and branch/model of the network in a similar way to graph database and used graph matching to detect anomalies in power flows.

2.3 Graph Similarity, Matching and Recommender Algorithms in Digital Information

As part of the investigation to find a novel and more flexible representation of the Network Codes, we also consider an exploration of graph similarity in digital information and specifically how this is applied on graph representations of textual information. In many applications it is common to see objects modelled as graphs which are discrete structures with vertices and edges (Nkgau and Anderson, 2017). There are three generalised methods of modelling objects this way. These are *Graph Matching*, a method of finding a correspondence between the nodes and edges of two graphs (CONTE et al., 2004), *Graph Similarity*, a method of calculating a score to calculate how similar two graphs are (Zager and Verghese, 2008a) and *Recommender Algorithms*, representing a collection of content-based, collaborative filtering and graph based techniques that are commonplace in e-commerce and social networks (Wang et al., 2010). Aside from these application domains, similarity, matching and recommender algorithms have had a large volume of research and applicable outputs in the areas of biology, chemistry, computer networks and image processing.

Graph Matching otherwise uses concepts such as Edit Distance and Graph Isomorphism, contain several techniques based on enumeration (Harary and Palmer, 1973) and approximation (Johnson, 1974) which have varying levels of suitability based on scalability, in terms of the size of the domain and also in terms of the combinatorial complexity of the problem. Common applications, among others, for Graph Matching include

image similarity , where a graph derived from an image is compared to a databases of graphs of known images (Bunke, 1997);

Semantic Matching , which uses an associated ontology to define the possible concepts, and matches the vertices and edges of the graph-based display of information to create a semantic graph of the data Cinque et al. (1996),

Structural Matching as defined by Ullmann (1976), operates on the depth-first tree search that specifies all potential mappings between the vertices on Graph A to those in Graph B, and the item[Similarity-Based Matching] approach which relies on using distance metrics like Euclidean distance (Bunke, 1999) to rank the similarity between objects in two graphs.

While advances have been made through decades of research the issue of exponential complexity remains and to mitigate against this combinational constraints are applied (Zhou and De la Torre, 2016) but by the application of these constraints the analysis could be susceptible to local optima which may exclude valid matchings contained in the wider dataset (Barbulescu et al., 2000).

Graph Similarity is based on two main techniques:

feature extraction , which operates on the idea that similar graphs share certain properties (Watts, 2004) and a group of techniques that centre on

iterative methods , which use the notion that 'two nodes are similar if their neighbourhoods are also similar'.

Feature extraction is the first step in a process and its purpose is to simplify the data in order to apply similarity measures (Cha, 2007), like Jaccard Coefficient (Niwattanakul et al., 2013) or Neighbour Matching (Nikolić, 2012). According to Koutra et al. (2011), feature extraction is powerful and scalable due to it breaking down the graph into statistical measures which are much smaller and easier to traverse than the input graphs. However feature extraction depends on the statistical measure chosen to provide an acceptable level of intuitiveness in the results. Iterative methods operate over the graph and each node is analysed and scored on similarity until all nodes have been analysed for similarity with all other nodes (Koutra et al., 2011). Iterative methods apply a category of algorithms, like similarity flooding algorithms (Melnik et al., 2002) and SimRank (Jeh and Widom, 2002) methods. Several works by, Cason et al. (2013); Koutra et al. (2011); Zager and Verghese (2008b) (among others) assess and measure the success, scalability and suitability of such techniques and algorithms and a common theme throughout is that the structure, size and complexity of the input graphs all have a bearing on how each technique is used and optimised. This iterative process is centred on the flow of choosing the candidate features, assessing that choice, re-factoring the criteria that the choice is subject to and choosing the candidate features again.

Recommender Algorithms, as stated earlier, are based on three key concepts, content-based recommendation, collaborative filtering and knowledge graph-based recommendation. Content-based recommendation is a technique which recommends resources based on their content and not on user's rating and opinion (Ferman et al., 2002) and according to Wang et al. (2010) has the disadvantage of needing structured resources, rich metadata and the taste of users should be captured in the features of the content. Collaborative filtering is based on the idea that similar users have similar interests and can be divided into two subcategories, user based (Harpale and Yang, 2008) and item based (Chen et al., 2008) and requires a user-item rating matrix similar to that proposed by Gao et al. (2017). It centres on finding similarity between the nearest neighbours and cosine-based similarity is often used (Wang et al., 2006). Knowledge graph-based methods try to learn effective representations of users and items according to the user-item interaction graphs and item-entity knowledge graphs, and then match the items to the users according to learned representations (Song et al., 2019).

While the above tools and techniques are mentioned and researched extensively in the context of social networks, biomedical, chemical, computer networks and image processing, examples of the use of such techniques in the context of electrical systems are quite new. There are examples of the use of Graph Similarity and Matching in system representation analysis (Santodomingo et al., 2014), security threat detection (Rawat and Bajracharya, 2015), household energy usage profiling (Charlton et al., 2013; Li and Dick, 2019) and power flow anomaly detection (Anwar and Mahmood, 2016).

2.4 Summary of Literature Review Findings

The previous section investigates the current state of the art from the context of how traditional documents are digitalised, how digital knowledge is represented in the electrical supply industry today and the processes, techniques and algorithms currently being applied in graph representations of textual representation. It was identified that there is an obvious gap in terms of how regulatory documents are represented in the smart grid in comparison to other sectors like legal and biomedical. This gap is in clear contrast with the advancements and research in the digitalisation of the electrical grid in terms of grid representation with Geographical Information System (GIS) and Common Information Model (CIM), for example, and JavaScript Object Notation (JSON) for data exchange. Currently the representation of the network codes require interpretation from the expert to enact, monitor or manually input their rules and constraints in the advanced automated digital systems that are becoming common place in electrical systems. The regulatory documents are a function of the grid, as identified in section 1.2.2, and it could be argued that they should be derived from some model of the entire electrical system but at present the electrical grid is represented in several different, non federated representations, often in different formats. CIM is emerging as a representation that could, in the future, provide this federated model, the current state does not allow for the extraction of the regulatory constraints in a comprehensive way. Table 2.1 puts the levels of digitalisation of some of the systems in the context of the Network Codes under three key aspects that are core to this dissertation, are they queryable, how are they exchanged and can a computer read them. On examining this table it is clear, particularly in the Exchange Format column, that the current representation of the Network Codes is not at the same level of digitalisation as some of the other key systems. It must be noted, however, that while the systems used for realtime monitoring and operations purposes cannot be compared directly to policy and compliance documents the difference in the levels of digitalisation between both groupings might inhibit future use cases where the compliance and policy documents may be required as input to an automated system. While there have been significant advancements in the area of text mining and digital similarity techniques it is not clear if these advancements have been made applicable to regulatory documents to the same extent as, for example, social media behaviour. The literature indicates that some work has been done in the legal sector but it has not been applied in the area of electrical regulatory documents as a use case.

Item	Queryable	Exchange Format	Computer Readability
Geographical Information System	Yes, normally via a bespoke user interfaces or database queries	XML, JSON or CSV	Data stored is computer readable with user interfaces to display information to the user
Meter Data Management System	Yes, via dashboards	JSON or CSV	Yes, the data is received from the meters and stored
Supervisory Control and Data Acquisition	Yes, via monitoring dashboards and database queries	JSON, CIM and CSV	The control messages and telemetry are received and stored.
Outage Management Systems	Yes, the alerts can be queried in realtime and historically	JSON, XML	The systems can detect a fault, react and store that fault information for querying later
Network Codes	Yes, using a find command in a document editor	PDF or Text Files	A computer can perform a basic search and can render the document but is not readable in the true sense without some preprocessing

Table 2.1 Summary of Digitalisation in the Smart Grid

Chapter 3

Research Context and Scope

3.1 Hypothesis

Network Codes can easily diverge, as highlighted in section 1.1, so there is a need to ensure that a semantically aligned, traceable digitised version of the Network Codes should be available to ensure their relevance and suitability in modern grid systems.

Formally, the research hypothesis is presented as:

"A set of technological processes using text mining and an enhanced digital representation of the Network Codes will help provide a more semantically aware, digitised and interactive version of the network codes.“

3.2 Research Questions

In the Literature we discussed the state of the art in how documents are digitised, how digital knowledge is represented in the smart grid, how semantic alignment is achieved in smart contracts and the current state of the art in graph similarity in digital information and we have noted gaps where the state of the art in these areas does not the address the research questions below. While the Network Codes are in a digital format they are in format that is primarily centred on human readability and lack features required for use as a knowledge representation in the Smart Grid that offers automated semantic matching. In other fields, like law and biomedical devices, this area has been advanced and large bodies of work and extensive parallel corpora have been amassed to aid the building of such systems and to form a layer of validation for newly digitised documents. As of yet this research has not been carried out on Network Codes, there are no corpora available and most interactions with the network codes are primarily of the manual kind. In the literature review section on Digital Knowledge Representation in the Electrical Supply Industry it was noted that there is a shift towards digitalisation across the industry and as discussed, the network codes are not in a format, at present to be capable of contributing and being a useful digital resource in the new digital electrical supply industry. This lack of a digital representation of multiple network codes is a gap in the State of the Art in how these documents are represented and shows a clear divergence in the level of digitalisation between how they are represented, compared to other data sources in the electrical grid. It is with this divergence in mind and to define the scope of this research, that two research questions, RQ1 and RQ2, have been proposed, with each question divided into sub questions as follows.

- RQ1: How do we represent the data from a Network Code document so that a subset can be used?
  - RQ1a: How do we label key concepts in the data?
  - RQ1b: How do we create a schema to represent the key concepts?
  - RQ1c: What impact does a change in an item in the annex have on the entire set of network codes?
- RQ2: Given there may be more than one set of Network Codes, how do we find commonalities across them?
  - RQ2a: What types of relationships between the Network Codes do we need to support?
  - RQ2b: Can we define a process to map network codes to each other with precision

3.3 Scenarios

To frame these research questions in a real world setting, two scenarios have been defined and it is these scenarios that will link the research, the experiments and the results to the real world problems identified in the Challenge and provide a grounding for this research in the context of how the The Electrical Industry has/is evolving from an analog, primarily copper based system towards full digitalisation. So it is with this in mind that two real world scenarios are presented and if we look at *Scenario 1*, which is based on the problems faced by network planners at present, and *Scenario 2*, which relates more to future requirements that are becoming more common.

3.3.1 Scenario 1: An electrical engineer wants to view all references to European Standard EN50160 and to assess the impact of a change across the document

In the ESB Networks Distribution Code Version 5.0 there is reference throughout the document to an European Standard called EN50160 EN50160 (1999), which is a standard that defines the main characteristics of the voltage at a network user's supply terminals in public low voltage and medium voltage electricity distribution systems under normal operating conditions. The reference to this standard is contained in the Supplementary Publications area of Annex 1 indicating that the actual text of the standard is held elsewhere in either the ESB repository or somewhere else that can be looked up using the annex reference. What if there were a need to replace this standard with a new one, for example EN50160? How would an expert assess the impact of such changes across the grid code document, policy and standard documents? One way would be to search for all instances of the code within the document using the `find all` function offered by the document viewer application, but unless the expert is extremely proficient in using the find function, potential references (with characteristics that do not match the specific search terms used) can be missed and furthermore the expert can only view one reference at a time. So to formalise the scenario, a fictitious change is to be made to how EN50160 is being referenced in the document. This involves all references to be changed to "EN50160-B" with the scenario assumption being that the old standard has been extended to accommodate prosumer interaction at the LV and MV user supply terminals. The expert now needs to assess whether it is better to appending the new requirements to an existing code or to create a new one. To do this, the expert will need to assess where, across the entire document, these schedules are used and also to identify if there are other codes referenced within the identified codes to assess the wider impact.

3.3.2 Scenario 2: An electrical engineer wants compare the codes in two sets of grid code documents

In the ESB Networks Distribution Code Version 5.0 and in the Eirgrid version there are codes duplicated in both documents and there are some unique to both sets that are specific to the DSO and the TSO respectively. The duplicated codes might be represented exactly, as a component of another code or even be worded slightly differently but with the same sentiment. What if there were regulatory changes being introduced that had an impact on both sets of codes and the changes were required to be to be mirrored across both sets. How would electrical engineers approach firstly assessing the changes and then making the changes consistently throughout both sets of codes? Would they search manually for the similar code representations independently and then make the changes with each organisation being responsible for their set? Would they work together to identify the matching codes and make the changes in tandem? Both these processes are feasible and with expert knowledge have a high degree or precision but have the potential to be slow and subject to human error. So to present this scenario in the context of both sets of grid codes, a fictitious change to the regulations regarding how data is handled by system operators is to be made across all grid codes in the EU to add an extra layer of governance around what data should be made available and how meaning the grid codes must be specific as to the data they reference in each code. The impact of this change is unknown within each set of regulatory documents but the first task set out by the regulator is for each system operator to identify the relevant codes that pertain to data so that a coherent approach can be developed around how these changes are going to be implemented. To do this, the experts will need to identify the relevant codes and collaborate to identify the matching codes so that any modifications would be reflected accurately.

Chapter 4

Research Methodology

In the Scenarios section we detailed two scenarios that frame the real-world context of how the research questions are answered. Based on these scenarios and the research questions, two experiments have been carried out.

4.1 Scenario 1 - Assessing the impact of Distribution Code changes

4.1.1 Introduction

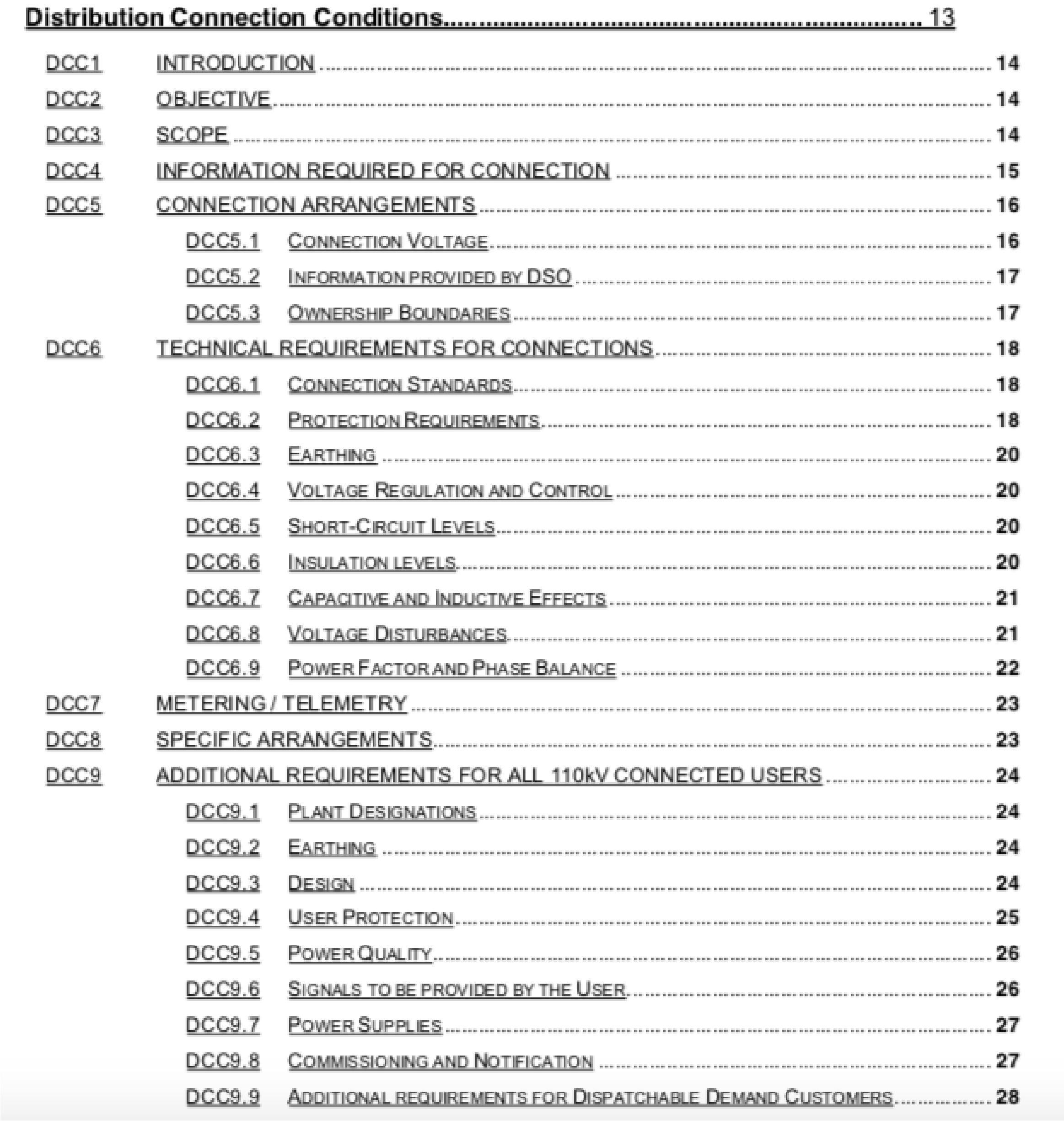
In order to address the impact of changes, see Section 3.3.1, we must look at it in the context of the research questions and sub questions relevant to it. Research question, How do we represent the data from a Network Code document so that a subset can be used?, depends upon the representation of the network codes and at technical level it involves taking them from a textual based document into something that can address the needs of the scenario. The needs of the scenario include being able to assess the impact of the change to a particular code across the entire document. In their current form this assessment involves the use of a find and replace for specific text across the document, which is a perfectly acceptable approach, but if the task became more complex, where there were other documents, or a multi search term scenario needs to be addressed this task could potentially become more complex. To format the content of the document in a way that supports search term flexibility and the extraction of a subset of the codes based on the properties of the text to achieve the goals of the scenario. To answer this, it is necessary to answer the research questions below.

- RQ1a: How do we label key concepts in the data?
- RQ1b: How do we create a schema to represent the key concepts?
- RQ2c: What impact does a change in an item in the annex have on the entire set of network codes?

The following work contains a set of experiments and defines the process used to answer the questions above. This will be done by constructing a set of digitalised network codes that will support the queries needed for Scenario 1.

4.1.2 Experiment 1 - Key Concept Labelling

After the ingestion phase where the provenance of the documents were verified by extracting and viewing the meta-data and the text of the document was converted from PDF formatted text to raw text, in order to enable a system to traverse this text in a meaningful way, key concepts needed to be identified and labelled. This experiment involves identifying the key concepts in the data, labelling them appropriately so that they can be traversed and thus acting as a *feature extraction* step as part of the graph similarity process identified in the literature review item on Graph Similarity, Matching and Recommender Algorithms in Digital Information that will be used in section 4.2.3, identified and assessed in isolation from the raw text that was parsed in Experiment 1. To achieve this it is essential to first visually inspect the document so that we can get a sense of the structure of the text, the properties of the document and an example of the concepts and their individual structure. An initial inspection yielded the presence of a Table of Contents and in figure 4.1, which is a sub set of the entire table of contents, it is shown that each section has a label, a heading and a page number associated with it. It was initially thought that the table of contents, would provide a starting place for the development of the logic that would extract the concepts from the raw text and map them to their parent concept. After initial experiments it was proven that the table of contents did not provide enough precision or granularity in terms of the codes and headings to capture all sections. This was evident when examining the example of *DCC5 CONNECTION ARRANGEMENTS* which has 3 child concepts, DCC5.1, DCC5.2 and DCC5.3, these are clearly shown in the table of contents but if the codes are examined it is evident that there subsections below each child. This is shown in figure 4.7 and while initial attempts were made to tokenise them using the headings and codes in the table of contents and detect the level of nesting of subsections were somewhat successful, that approach was found to be brittle due to table of contents codes and headings being repeated throughout the document in an inconsistent way leading to complex regular expressions required to extract the concepts in a meaningful way. While the table of contents provided a high level pointer to where sections were it did not provide a representation of the true pattern of the document. Thus it was decided to move away from the approach of using the table of contents to drive the extraction of the concepts in favour of an approach that would examine the document in a less informed way in terms of section nesting and chapter structure and allow the system to detect the nesting in an autonomous way.



Distribution Connection Conditions 13

DCC1 INTRODUCTION 14
DCC2 OBJECTIVE 14
DCC3 SCOPE 14
DCC4 INFORMATION REQUIRED FOR CONNECTION 15
DCC5 CONNECTION ARRANGEMENTS 16
DCC5.1 CONNECTION VOLTAGE 16
DCC5.2 INFORMATION PROVIDED BY DSO 17
DCC5.3 OWNERSHIP BOUNDARIES 17
DCC6 TECHNICAL REQUIREMENTS FOR CONNECTIONS 18
DCC6.1 CONNECTION STANDARDS 18
DCC6.2 PROTECTION REQUIREMENTS 18
DCC6.3 EARTHING 20
DCC6.4 VOLTAGE REGULATION AND CONTROL 20
DCC6.5 SHORT-CIRCUIT LEVELS 20
DCC6.6 INSULATION LEVELS 20
DCC6.7 CAPACITIVE AND INDUCTIVE EFFECTS 21
DCC6.8 VOLTAGE DISTURBANCES 21
DCC6.9 POWER FACTOR AND PHASE BALANCE 22
DCC7 METERING / TELEMETRY 23
DCC8 SPECIFIC ARRANGEMENTS 23
DCC9 ADDITIONAL REQUIREMENTS FOR ALL 110kV CONNECTED USERS 24
DCC9.1 PLANT DESIGNATIONS 24
DCC9.2 EARTHING 24
DCC9.3 DESIGN 24
DCC9.4 USER PROTECTION 25
DCC9.5 POWER QUALITY 26
DCC9.6 SIGNALS TO BE PROVIDED BY THE USER 26
DCC9.7 POWER SUPPLIES 27
DCC9.8 COMMISSIONING AND NOTIFICATION 27
DCC9.9 ADDITIONAL REQUIREMENTS FOR DISPATCHABLE DEMAND CUSTOMERS 28

Fig. 4.1 Distribution Code Table of Contents Example.

The new approach involved a three step process that uses basic terms to categorise the codes initially and to extract the text and structure of the codes. This three step process requires a set of traversals across the raw text to first clean, then segment and lastly define the nesting structure of the sections. Figure 4.2 represents the process that will extract the concepts from the raw text while maintaining the structure of the subsections and their nesting. The following details the three stages process to extract a full representation of *DCC5.1 Connection Voltage* from the raw text.

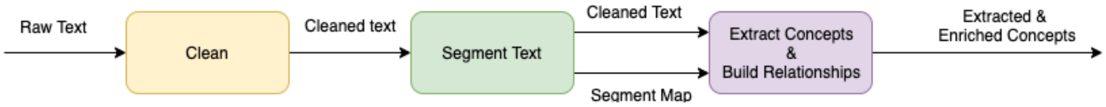

Raw Text
Clean
Cleaned text
Segment Text
Cleaned Text
Segment Map
Extract Concepts & Build Relationships
Extracted & Enriched Concepts

Fig. 4.2 Distribution Code Concept Extraction Process.

In order to eliminate unwanted content, like the table of contents, punctuation and footer content, it was important to clean the data so that the cleanest possible representation of the raw data was passed to the segmentation process. Figure 4.3 illustrates these steps. These were carried out using regular expressions that identified the items and either removed them or replaced them with a suitable identifier. The regular expressions were created manually and iteratively re-factored throughout the development process, if there was a corpus present and expert input it could be conceivable that these regular expressions could be populated with verified keywords making them transferable to other documents with the same subject matter. The resulting data, at this stage should be suitably cleaned and correctly formatted to be iterated across in the Grid code segmentation process in figure 4.4.

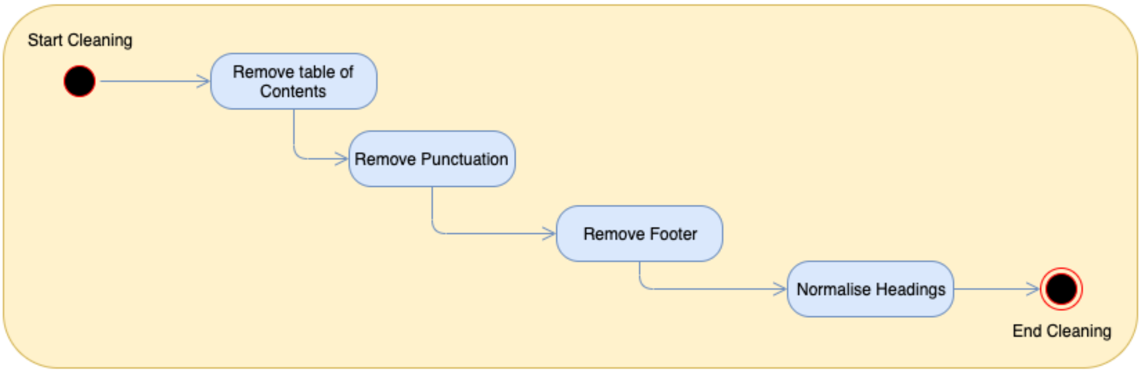


Fig. 4.3 Grid code cleaning process

Given that the distribution codes have a defined set of sections but within these sections there are repeating terms, like in the Preface section there are references to codes and terms within other sections, it was decided to make a logical map of these segments in order to treat each segment in isolation when extracting the concepts. Given that the headings are predefined it was decided to feed the segmentation process with a map of the key terms that define each segment of the document. This map was used to identify the start and end of the each section when iterating through the data. Over the course of iterating over the data a mapping was built up that will guide and strengthen the process of extracting the concepts. Figure 4.5 is the output of the segmentation process and, as can be seen, it normalises the names and crucially contains a start and end value which denote the line count at the start and end of each segment. It is this segment map, along with the cleaned data, that is fed to the concept extraction process.

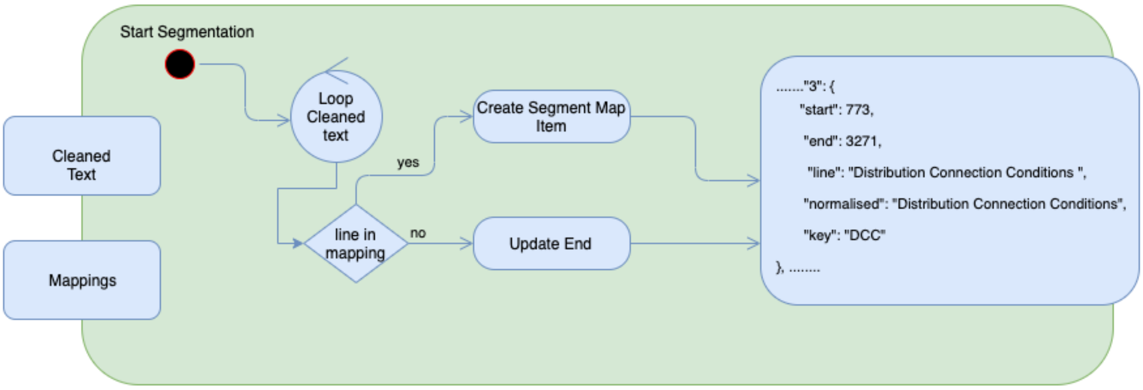
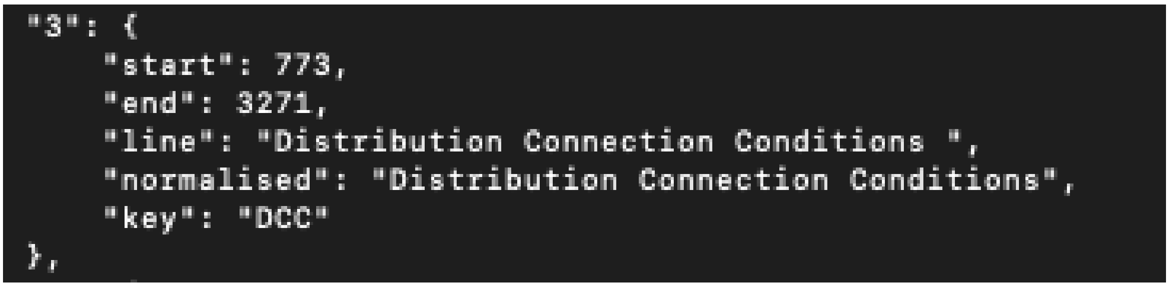


Fig. 4.4 Grid code segmentation process



```
"3": {
    "start": 773,
    "end": 3271,
    "line": "Distribution Connection Conditions ",
    "normalised": "Distribution Connection Conditions",
    "key": "DCC"
},
```

Fig. 4.5 Grid code segments

The extraction process is logically the most complex of the stages of this experiment and as can be seen from figure 4.6 it involves using the details in the mapping to extract the separate concepts and build up the relationships to maintain the structure of the document regarding sections and the nested structure.

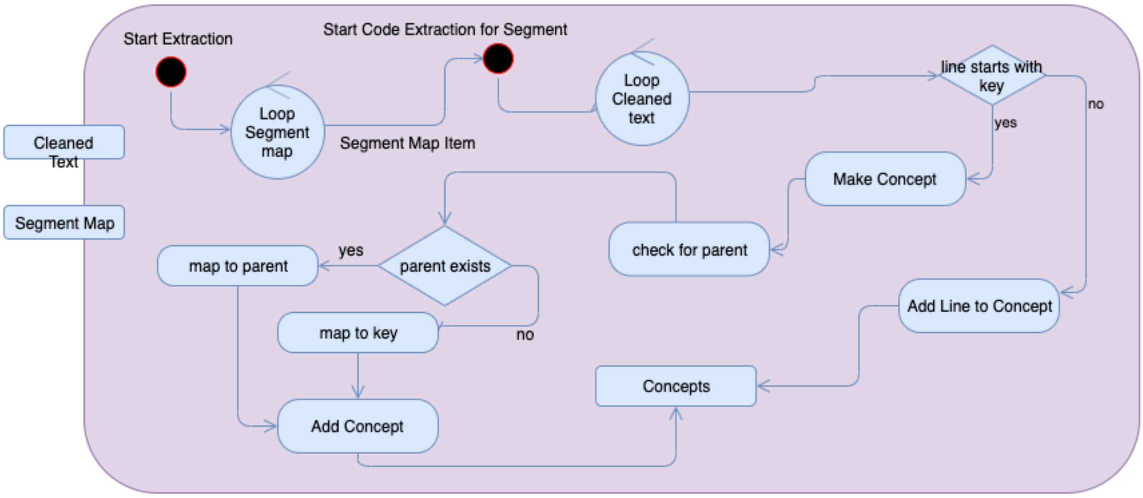
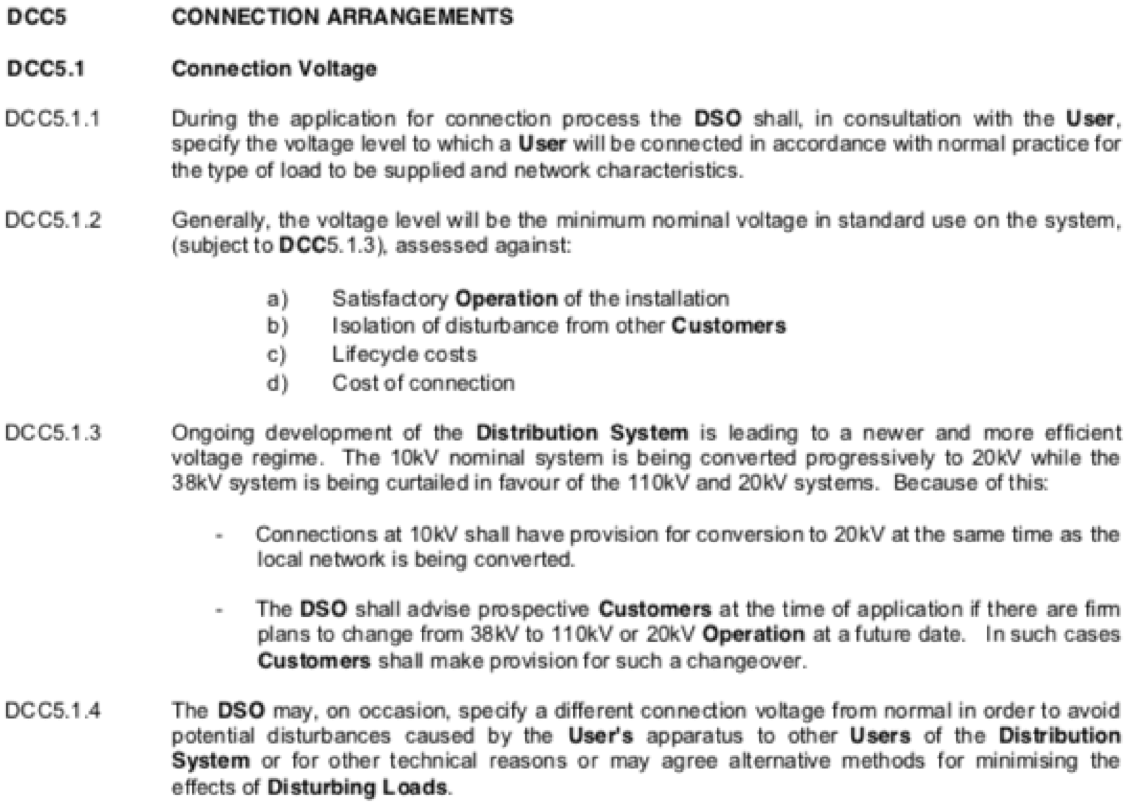


Fig. 4.6 Grid code extraction process

4.1.2.1 Result

The main goal of this experiment was to label the concepts in such a way that the concepts were clearly defined and that the features were extracted, the structure of the document was maintained in logic and that the relationships contained between the codes were preserved so that the scope of use of the data was not hindered. An initial attempt was made using the table of contents as a driver for the extraction of the concepts but that was found not to be viable due to a lack of precision and depth in the content of the table of contents. This approach was changed in favour of the three step process presented above. There were three phases in this experiment that needed to be carried out sequentially with the output and the quality of that output vital to the success of the next task in the process. As can be seen from figures 4.7 and 4.8, a before and after representation of DCC5 and its subsections, the system extracted the text and the subsections sufficiently. Each of these represents a concept within the document and in the properties section contains the relationship details for each concept. This will allow the correct mappings to take place within the chosen schema and allow a broader set of use cases to be developed using both the content and the relationships.



DCC5 **CONNECTION ARRANGEMENTS**

DCC5.1 **Connection Voltage**

DCC5.1.1 During the application for connection process the **DSO** shall, in consultation with the **User**, specify the voltage level to which a **User** will be connected in accordance with normal practice for the type of load to be supplied and network characteristics.

DCC5.1.2 Generally, the voltage level will be the minimum nominal voltage in standard use on the system, (subject to **DCC**5.1.3), assessed against:

a) Satisfactory **Operation** of the installation
b) Isolation of disturbance from other **Customers**
c) Lifecycle costs
d) Cost of connection

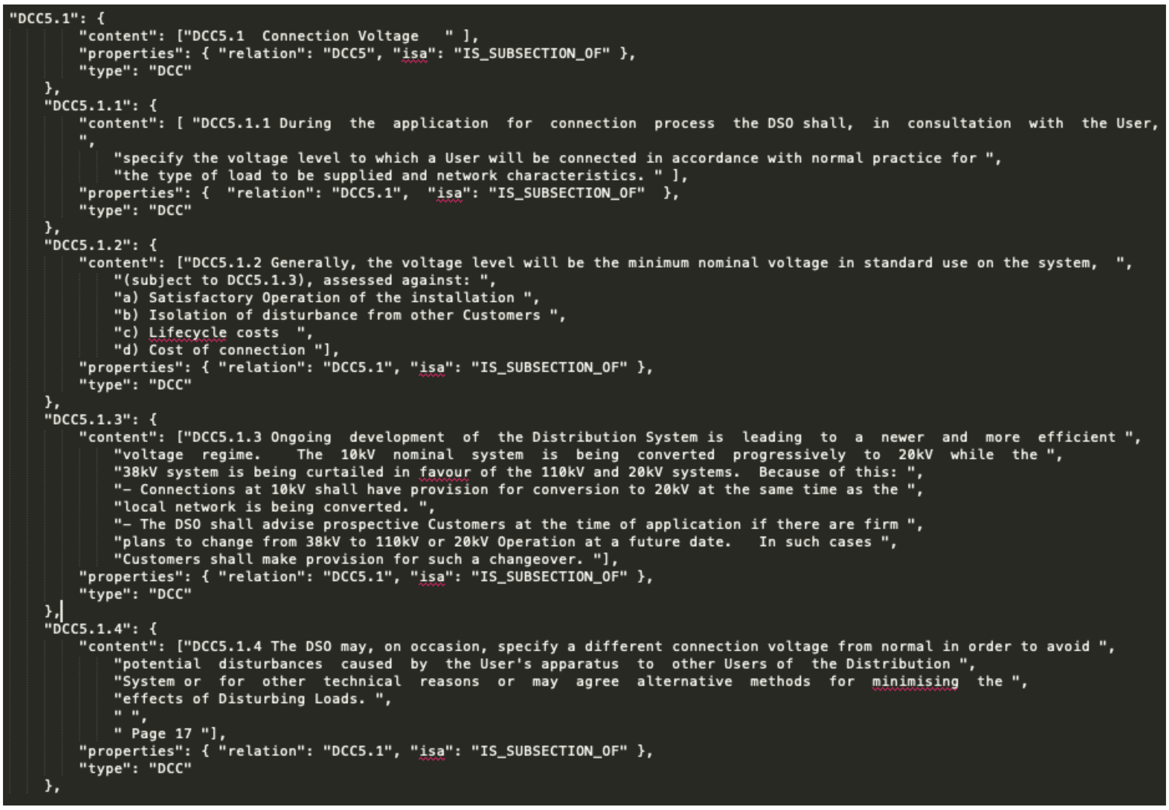
DCC5.1.3 Ongoing development of the **Distribution System** is leading to a newer and more efficient voltage regime. The 10kV nominal system is being converted progressively to 20kV while the 38kV system is being curtailed in favour of the 110kV and 20kV systems. Because of this:

- Connections at 10kV shall have provision for conversion to 20kV at the same time as the local network is being converted.

- The **DSO** shall advise prospective **Customers** at the time of application if there are firm plans to change from 38kV to 110kV or 20kV **Operation** at a future date. In such cases **Customers** shall make provision for such a changeover.

DCC5.1.4 The **DSO** may, on occasion, specify a different connection voltage from normal in order to avoid potential disturbances caused by the **User's** apparatus to other **Users** of the **Distribution System** or for other technical reasons or may agree alternative methods for minimising the effects of **Disturbing Loads**.

Fig. 4.7 Distribution Code DCC5



```
"DCC5.1": {
        "content": ["DCC5.1  Connection Voltage   " ],
        "properties": { "relation": "DCC5", "isa": "IS_SUBSECTION_OF" },
        "type": "DCC"
    },
    "DCC5.1.1": {
        "content": [ "DCC5.1.1 During  the  application  for  connection  process  the DSO shall,  in  consultation  with  the User,
        ",
            "specify the voltage level to which a User will be connected in accordance with normal practice for ",
            "the type of load to be supplied and network characteristics. " ],
        "properties": {  "relation": "DCC5.1",  "isa": "IS_SUBSECTION_OF"  },
        "type": "DCC"
    },
    "DCC5.1.2": {
        "content": ["DCC5.1.2 Generally, the voltage level will be the minimum nominal voltage in standard use on the system,  ",
            "(subject to DCC5.1.3), assessed against: ",
            "a) Satisfactory Operation of the installation ",
            "b) Isolation of disturbance from other Customers ",
            "c) Lifecycle costs  ",
            "d) Cost of connection "],
        "properties": { "relation": "DCC5.1", "isa": "IS_SUBSECTION_OF" },
        "type": "DCC"
    },
    "DCC5.1.3": {
        "content": ["DCC5.1.3 Ongoing  development  of  the Distribution System is  leading  to  a  newer  and  more  efficient ",
            "voltage  regime.   The  10kV  nominal  system  is  being  converted  progressively  to  20kV  while  the ",
            "38kV system is being curtailed in favour of the 110kV and 20kV systems.  Because of this: ",
            "- Connections at 10kV shall have provision for conversion to 20kV at the same time as the ",
            "local network is being converted. ",
            "- The DSO shall advise prospective Customers at the time of application if there are firm ",
            "plans to change from 38kV to 110kV or 20kV Operation at a future date.   In such cases ",
            "Customers shall make provision for such a changeover. "],
        "properties": { "relation": "DCC5.1", "isa": "IS_SUBSECTION_OF" },
        "type": "DCC"
    },
    "DCC5.1.4": {
        "content": ["DCC5.1.4 The DSO may, on occasion, specify a different connection voltage from normal in order to avoid ",
            "potential  disturbances  caused  by  the User's apparatus  to  other Users of  the Distribution ",
            "System or  for  other  technical  reasons  or  may  agree  alternative  methods  for  minimising  the ",
            "effects of Disturbing Loads. ",
            " ",
            " Page 17 "],
        "properties": { "relation": "DCC5.1", "isa": "IS_SUBSECTION_OF" },
        "type": "DCC"
    },
```

Fig. 4.8 Distribution Code DCC5.1 Extracted

4.1.3 Experiment 2 - Creating A Schema

The subject matter for this work, the distribution codes and their counterparts, the network codes, have only ever been represented in a textual format and one that can only be analysed either visually or using the tools provided by a PDF viewer or word processing tool. What that means in a business context is that any new representation of these will open up new use cases and potential business process uses that were previously over looked due to their rigid format. With this in mind it is important that when we define a schema to represent the concepts that does not only provide a correct representation of the codes but one that also is not overly restrictive as to limit the scope of future use cases or ways that business processes can be improved. This experiment is centred on choosing an object store and a schema to store these concepts in a way that will do both. Taking *DCC5 CONNECTION ARRANGEMENTS* as an example, we will discuss the options that were discovered, what was chosen, the result in the context of DCC5 and also present a brief example of how the schema chosen can be extended to cater for a new use case.

The three options initially explored as potential object stores for the concepts were *relational*, *non relational* and *graph* all of which are perfectly viable solutions but each has properties that lead to the creation of schemas that have varying characteristics and are suited to certain types of use cases. A relational type database was initially explored and it was found that any schema created here, while it would provide a rigid structure for the concepts, it was felt that is would be too restrictive in terms of the definition of column sizes and the enforcement of relationships, it was also felt that any future use cases would need to be analysed deeply prior to the creation of a relational schema. At this stage this is not possible due to the novel nature of the representation of the codes. Given that the output of experiment 1, presented in figure 4.8, is in JSON Crockford (2006) format the obvious choice seemed to be a Mongodb object store that can take the concepts extracted from the codes and create collections and documents from them. Using the information in the properties fields it might be possible to define the relationships between codes. While a non relational object store (such as the mongodb document database) was a viable solution and the representation of the data would mirror the concept extracted, to query the concepts with any level of complexity involving relationships would require complex queries. Indeed that would mean that the interrelationships between concepts would need to be queried explicitly (the query author would need a deep understanding of the structure of each document). The next object store that was explored was a graph database which treats the data as nodes and edges which can be mapped to the concepts and relationships extracted from the documents. The exploration of the use of graph databases as a method of representing the Network Codes was motivated by the recent research in the use of graph databases identified in Section 2.2. This solution treats the relationship as first-class entity within the object store that can be queried and one that can be built upon to create new relationships between concepts which would enable the schema maintain the structure of the document and also provide a mechanism to enable the new relationships be created between concepts that do not impact existing relationships.

Based on this initial investigation, it was decided to choose a Graph Database management system, Neo4J, as the object store and the concept extraction process presented in figure 4.2, was extended to store the concepts therein. Neo4J enables the modelling of a domain as a graph with each node containing a set of properties with the properties not directly dependant on the relationships the node participates in. This allows dynamic relationships to exist between nodes and some but not all of the properties. Figure 4.9 is an illustration of the process that was used to convert the concepts extracted into nodes with the content field storing the description of the node and the key forming the name of the node. The details in the properties field are mainly to drive the creation of the relationship between the nodes.

During the concept extraction phase of the previous experiment and from visually inspecting the document it was noted that there were two relationships that were required at the very least to maintain the structure of the document in a meaningful way and these are **IS_CATAGORISED_AS** and **IS_SUBSECTION_OF**. In the type field of the concepts, as presented in figure 4.8, the value here signifies that the code belongs to the category *Distribution Connection Conditions* and thus if a code has no parent code identified it is defaulted to the **IS_CATAGORISED_AS** relationship and related to a master node of that name. This will allow each node to have a relationship but will also enable the categorisation of concepts extracted from other entities to be related to concepts at the highest level at least. The **IS_SUBSECTION_OF** relationship is applied to the parent key contained in the relation field and this maintains the structure of the document as it will relate that node to the appropriate node at the next hierarchical level. Fully aware of the Neo4J specific Labelling system, could be used to help with categorisation, the pure property graph approach was preferred for the purpose of interoperability making the potential export to an RDF model easier minus the need for Neo4J extensions.

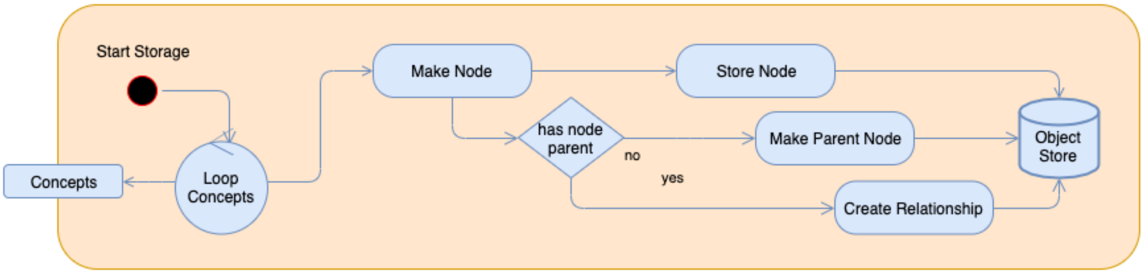


Fig. 4.9 Grid Code Storage Process

4.1.3.1 Result

The results generated from the above experiment are presented via the web server provided by Neo4j which allows the querying of the nodes and edges and provides a graphical or textual representation of the data. As mentioned in the experiment section above we will maintain the use of *DCC5 CONNECTION ARRANGEMENTS* to present these results. To extract the concepts generated from DCC5 a query was written and executed to extract DCC5 and all its child concepts. The query was `MATCH (n:‘distributioncodev50‘) WHERE (n.name CONTAINS ’DCC5’) RETURN n` and figure 4.10 is the result of that query.

Note the structure of the nodes and relationships and, in comparison with figure 4.7, how the document structure is maintained. To further examine this we refine the query to look only at DCC5.1 and figure 4.11 is the results of that query presenting the content of the section and also the section code number. This is replicated across the entire document where there are 577 nodes created from the document with each node representing a code within the document and each node being related to a relevant node as either a category of or a subsection of another node in the database.

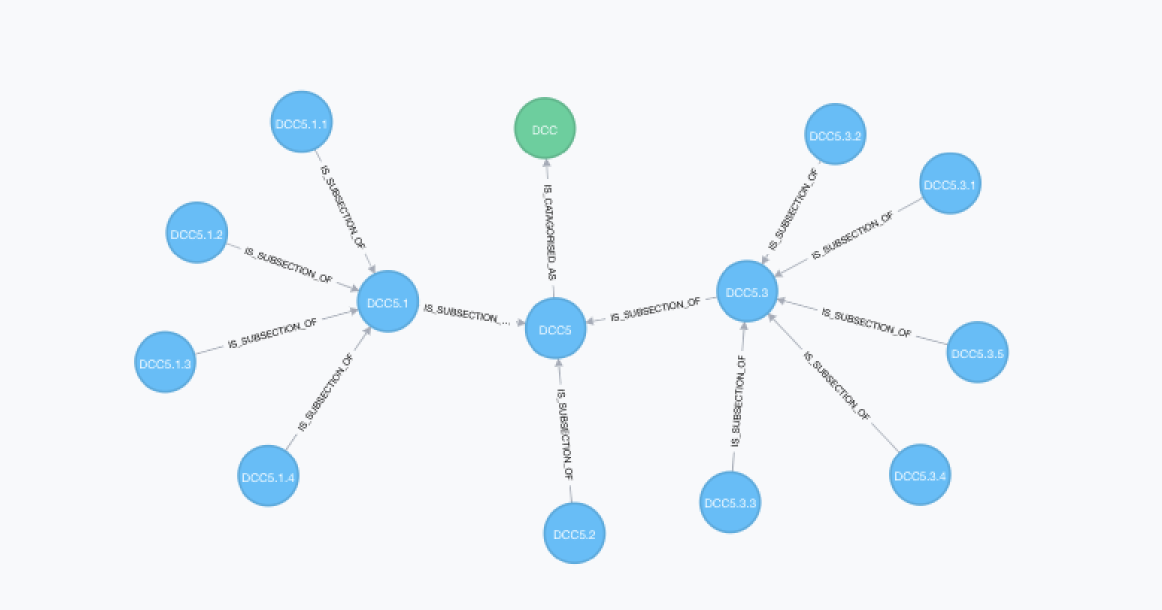
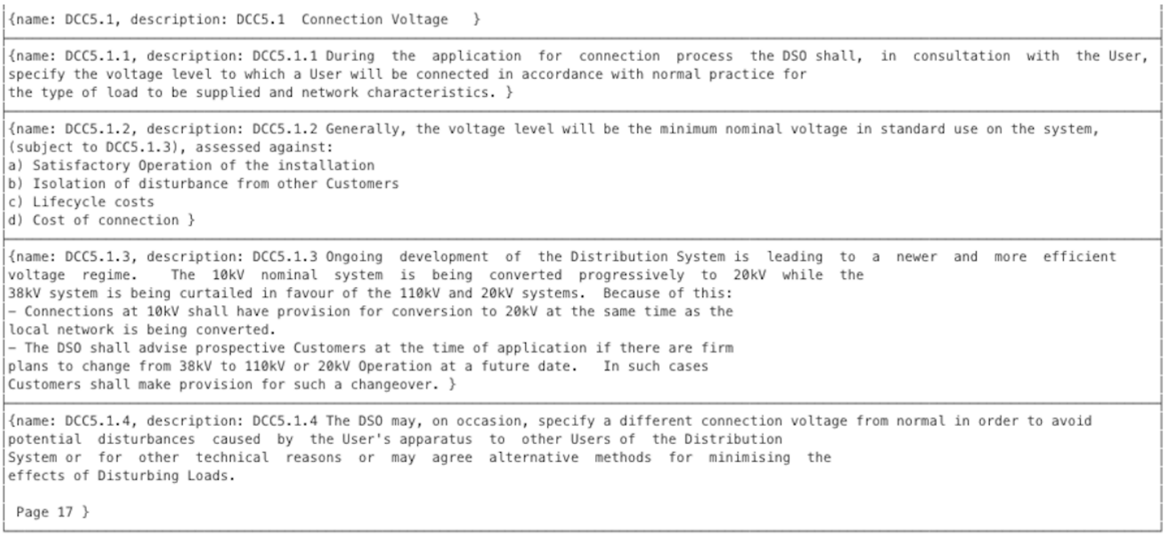


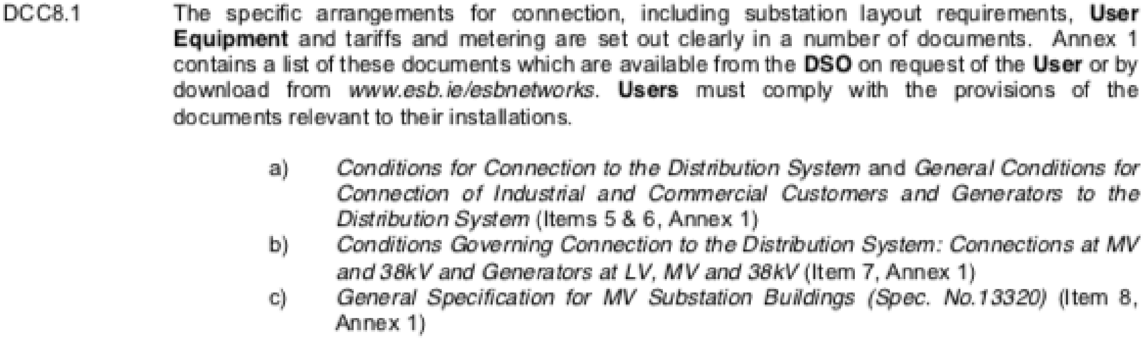
Fig. 4.10 DCC Concepts Stored in Neo4j



```
{name: DCC5.1, description: DCC5.1  Connection Voltage   }

{name: DCC5.1.1, description: DCC5.1.1 During  the  application  for  connection  process  the DSO shall,  in  consultation  with  the User,
specify the voltage level to which a User will be connected in accordance with normal practice for
the type of load to be supplied and network characteristics. }

{name: DCC5.1.2, description: DCC5.1.2 Generally, the voltage level will be the minimum nominal voltage in standard use on the system,
(subject to DCC5.1.3), assessed against:
a) Satisfactory Operation of the installation
b) Isolation of disturbance from other Customers
c) Lifecycle costs
d) Cost of connection }

{name: DCC5.1.3, description: DCC5.1.3 Ongoing  development  of  the Distribution System is  leading  to  a  newer  and  more  efficient
voltage  regime.   The  10kV  nominal  system  is  being  converted  progressively  to  20kV  while  the
38kV system is being curtailed in favour of the 110kV and 20kV systems.  Because of this:
- Connections at 10kV shall have provision for conversion to 20kV at the same time as the
local network is being converted.
- The DSO shall advise prospective Customers at the time of application if there are firm
plans to change from 38kV to 110kV or 20kV Operation at a future date.   In such cases
Customers shall make provision for such a changeover. }

{name: DCC5.1.4, description: DCC5.1.4 The DSO may, on occasion, specify a different connection voltage from normal in order to avoid
potential  disturbances  caused  by  the User's apparatus  to  other Users of  the Distribution
System or  for  other  technical  reasons  or  may  agree  alternative  methods  for  minimising  the
effects of Disturbing Loads.

 Page 17 }
```

Fig. 4.11 DCC5.1 Text and Name Storage

The nature of this type of database and the schema it provides is that new relationships can be created based on the properties of the text contained in the nodes. To demonstrate this we take Item 7 from Annex 1 from the network codes and we can create a relationship from that with all the other codes in the document. Item 7 covers the "Conditions Governing Connection to the Distribution System: Connections at MV and 38kV; and Generators at LV, MV and 38kV" and is referenced throughout the document. From doing a simple find operation using a PDF viewer it is shown to be present in the codes in 4 places, DCC10.2.2, DCC10.2.3, DCC6.3.2 and the Annex. Running an operation to match the term and create the relation between them and the Annex returns 4 nodes where a reference has been found excluding the Annex. This includes a code, DCC8.1, that has not been picked up in the find operation by the PDF viewer and on inspection it does contain reference to Item 7 in the Annex, see figure 4.12 item b. Figure 4.13 is the response to a query that will return and match Annex 1 with all references to Item 7 in the document and this reference was created and can be queried independently of all other relationships in the database which demonstrates the flexibility of the schema that would enable complex relationships be developed across the codes and other versions of the codes that do not impact on the underlying structure of the codes.



DCC8.1 The specific arrangements for connection, including substation layout requirements, **User Equipment** and tariffs and metering are set out clearly in a number of documents. Annex 1 contains a list of these documents which are available from the **DSO** on request of the **User** or by download from *www.esb.ie/esbnetworks*. **Users** must comply with the provisions of the documents relevant to their installations.

a) *Conditions for Connection to the Distribution System* and *General Conditions for Connection of Industrial and Commercial Customers and Generators to the Distribution System* (Items 5 & 6, Annex 1)

b) *Conditions Governing Connection to the Distribution System: Connections at MV and 38kV and Generators at LV, MV and 38kV* (Item 7, Annex 1)

c) *General Specification for MV Substation Buildings (Spec. No.13320)* (Item 8, Annex 1)

Fig. 4.12 DCC8.1 Text

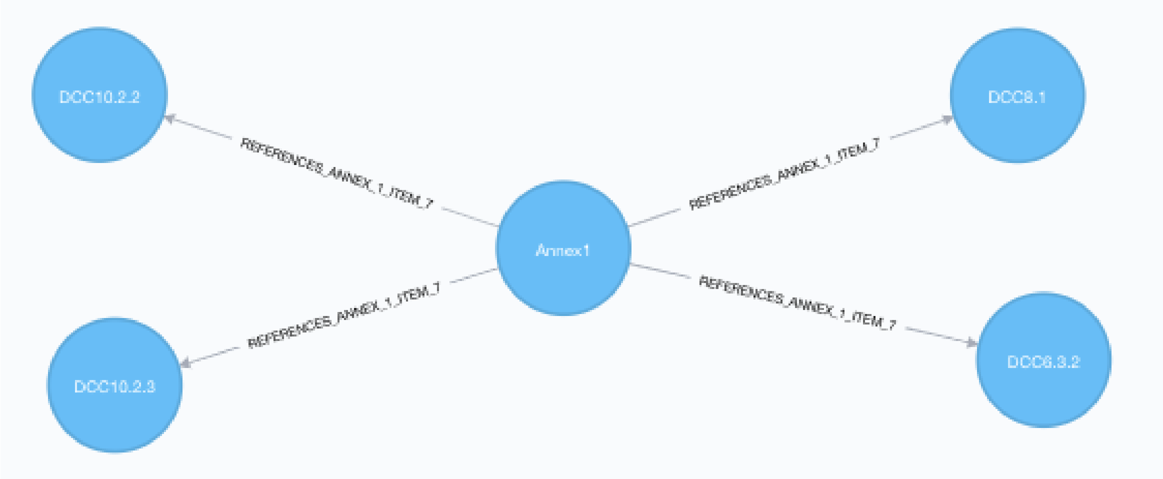
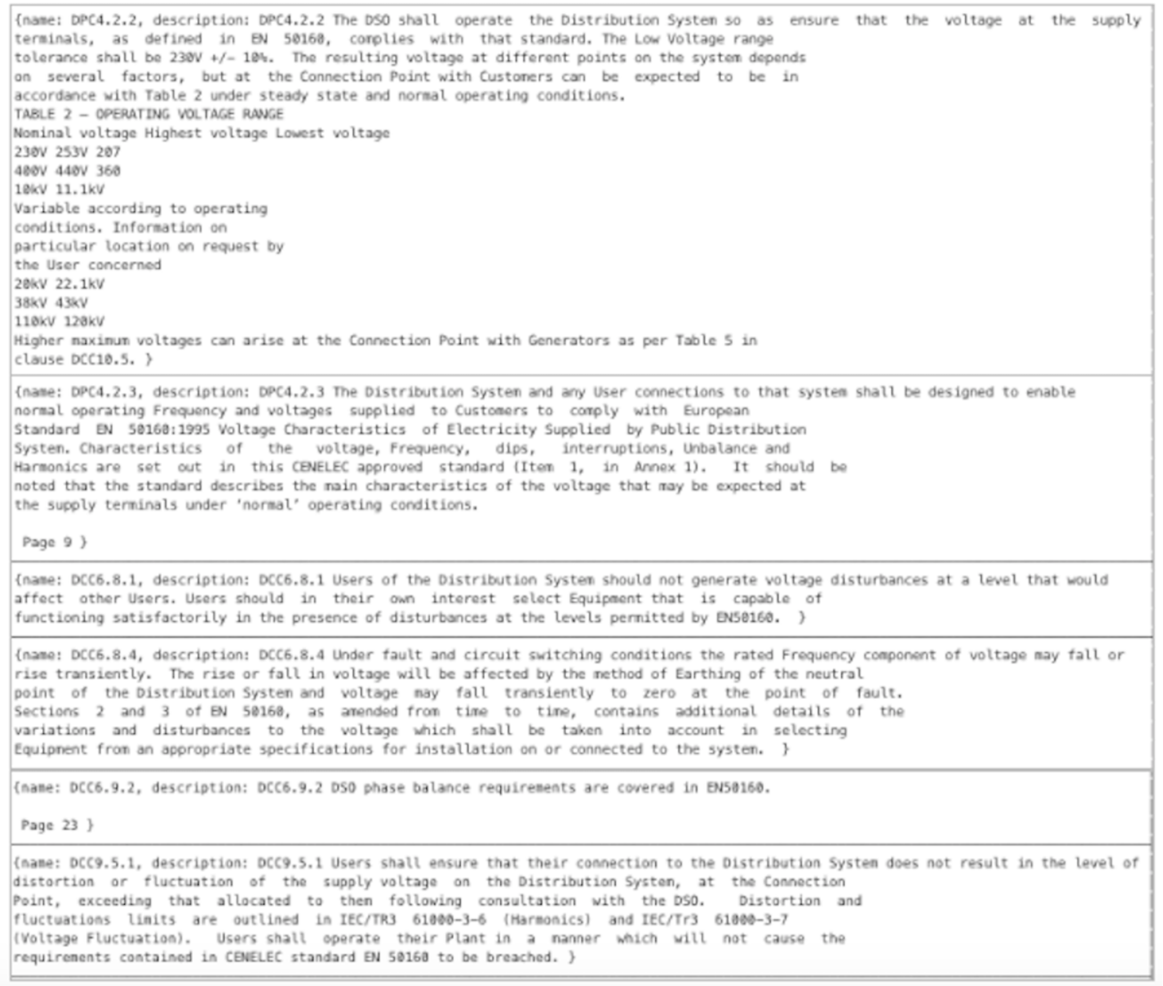

DCC10.2.2
DCC8.1
Annex1
DCC10.2.3
DCC6.3.2
REFERENCES_ANNEX_1_ITEM_7
REFERENCES_ANNEX_1_ITEM_7
REFERENCES_ANNEX_1_ITEM_7
REFERENCES_ANNEX_1_ITEM_7

Fig. 4.13 Annex 1 Item 7 document references

4.1.4 Experiment 3 - Assessing the impact of change

Using the concepts extracted and stored from the previous experiments the data is in a format where we can answer research question *RQ1d: What impact does a change in an item in the annex have on the entire set of network codes?* and to do this we first need to find a candidate Annex Item to assess. To achieve this we pose the fictitious scenario that a change in European Standard EN50160 (1999) was required and this change required that this standard be renamed to EN 50160b and some of the context of the standard was changed also. To assess this impact visually at present would require a find operation using a PDF viewer and as can be seen from the previous experiment some items can be missed using that approach due to search term misrepresentation. The assessment in this experiment will take the characters "50160" and the term "Item 1" as input into a query to gather all representations of these terms across the network codes where a visual inspection can take place to assess, first the context that each mention is used and also to examine the consistency of how it is represented throughout the document. To achieve this assessment a relationship could be created to match these mentions/references to the Annex but in this case it will be just carried out using a query. The query that will perform this is *MATCH (n:‘distributioncodev50‘) WHERE (n.description CONTAINS ’50160’ OR n.description CONTAINS ’Item 1,’) RETURN n* which will match all occurrences of the EN50160 by reference across the concepts stored in the database. The response when executing the query is presented in 4.14 and as can be seen it is referenced by 5 other codes as either Item 1 or as 50160.



```
{name: DPC4.2.2, description: DPC4.2.2 The DSO shall  operate  the Distribution System so  as  ensure  that  the  voltage  at  the  supply
terminals,  as  defined  in  EN  50160,  complies  with  that standard. The Low Voltage range
tolerance shall be 230V +/- 10%.  The resulting voltage at different points on the system depends
on  several  factors,  but at  the Connection Point with Customers can  be  expected  to  be  in
accordance with Table 2 under steady state and normal operating conditions.
TABLE 2 – OPERATING VOLTAGE RANGE
Nominal voltage Highest voltage Lowest voltage
230V 253V 207
400V 440V 360
10kV 11.1kV
Variable according to operating
conditions. Information on
particular location on request by
the User concerned
20kV 22.1kV
38kV 43kV
110kV 120kV
Higher maximum voltages can arise at the Connection Point with Generators as per Table 5 in
clause DCC10.5. }
```

```
{name: DPC4.2.3, description: DPC4.2.3 The Distribution System and any User connections to that system shall be designed to enable
normal operating Frequency and voltages  supplied  to Customers to  comply  with  European
Standard  EN  50160:1995 Voltage Characteristics  of Electricity Supplied  by Public Distribution
System. Characteristics   of   the   voltage, Frequency,   dips,   interruptions, Unbalance and
Harmonics are  set  out  in  this CENELEC approved  standard (Item  1,  in  Annex 1).   It  should  be
noted that the standard describes the main characteristics of the voltage that may be expected at
the supply terminals under 'normal' operating conditions.

 Page 9 }
```

```
{name: DCC6.8.1, description: DCC6.8.1 Users of the Distribution System should not generate voltage disturbances at a level that would
affect  other Users. Users should  in  their  own  interest  select Equipment that  is  capable  of
functioning satisfactorily in the presence of disturbances at the levels permitted by EN50160.  }
```

```
{name: DCC6.8.4, description: DCC6.8.4 Under fault and circuit switching conditions the rated Frequency component of voltage may fall or
rise transiently.  The rise or fall in voltage will be affected by the method of Earthing of the neutral
point  of  the Distribution System and  voltage  may  fall  transiently  to  zero  at  the  point  of  fault.
Sections  2  and  3  of EN  50160,  as  amended from  time  to  time,  contains  additional  details  of  the
variations  and  disturbances  to  the  voltage  which  shall  be  taken  into  account  in  selecting
Equipment from an appropriate specifications for installation on or connected to the system.  }
```

```
{name: DCC6.9.2, description: DCC6.9.2 DSO phase balance requirements are covered in EN50160.

 Page 23 }
```

```
{name: DCC9.5.1, description: DCC9.5.1 Users shall ensure that their connection to the Distribution System does not result in the level of
distortion  or  fluctuation  of  the  supply voltage  on  the Distribution System,  at  the Connection
Point,  exceeding  that  allocated  to  then  following  consultation  with  the DSO.   Distortion  and
fluctuations  limits  are  outlined  in IEC/TR3  61000-3-6  (Harmonics)  and IEC/Tr3  61000-3-7
(Voltage Fluctuation).   Users shall  operate  their Plant in  a  manner  which  will  not  cause  the
requirements contained in CENELEC standard EN 50160 to be breached. }
```

Fig. 4.14 Query Return for 50160

4.1.4.1 Result

There are many codes that reference 50160 in the document. Firstly, it is noted that there are some discrepancies concerning how the code is referenced. While a reference is held in Annex 1, Item 1, it is only cited throughout the document in one place, DPC4.2.3, *Characteristics of the voltage, Frequency, dips, interruptions, Unbalance and Harmonics are set out in this CENELEC approved standard (Item 1, in Annex 1). It should be noted that the standard describes the main characteristics of the voltage that may be expected at the supply terminals under 'normal' operating conditions.* elsewhere it is referenced by either "EN 50160" or "EN50160". This highlights a lack of alignment across the referencing of annex items across the document and given that the impact of such a change can be assessed given other changes, for example, a change to an item in the Distribution Data Registration schedules could be assessed across the network codes. In a meeting with Pádraig Lyons (an industry expert in energy policy documents)(Lyons et al., 2018) it was detailed that an assessment of the impact of a change involves a process whereby the expert would be provided with the change and then manually identify the codes where the change would need to be made. So to bring that into the concept of the experiment for the expert to assess the change, the expert would need to be aware that there are differences in how it is referenced at least to inform their search. Furthermore, the sample and the impact of the annex item chosen here is relatively small but if there was a more impactful change with a greater level of complexity the manual process would become more laborious and a user interface that would provide input to the queries defined in this experiment would provide the expert with potentially a less laborious method of identifying the codes impacted.

4.2 Scenario 2 - An electrical engineer wants compare the codes in two sets of grid code documents

4.2.1 Introduction

In order to address Scenario 2: An electrical engineer wants compare the codes in two sets of grid code documents, presented in section 3.3.2: the research question is: if there is more than one set of Network Codes, how do we find commonalities across them?. This research question and its sub questions explore the relationships that would be required across the multi-actor network codes and how, if it is possible, they can be derived from each representations. To answer these questions fully, by building a system that would provide a semantically aligned set of network codes, would require the implementation of techniques that are outside the scope of this research and would require full validation from industry in terms of the accuracy of the outcomes. Instead, this research has less ambitious goals: it shows how enhanced representations of network codes, and tools to support them, can make the task easier, and hints at possible future developments where semantic alignment can be achieved with greater automation.

- RQ2a: What types of relationships between the Network Codes do we need to support?
- RQ2b: Can we define a process to map network codes to each other with precision

The following experiments will outline the software steps and business logic that will answer the above with the fulfillment of Scenario 2 being the minimum requirement of the overall outcome.

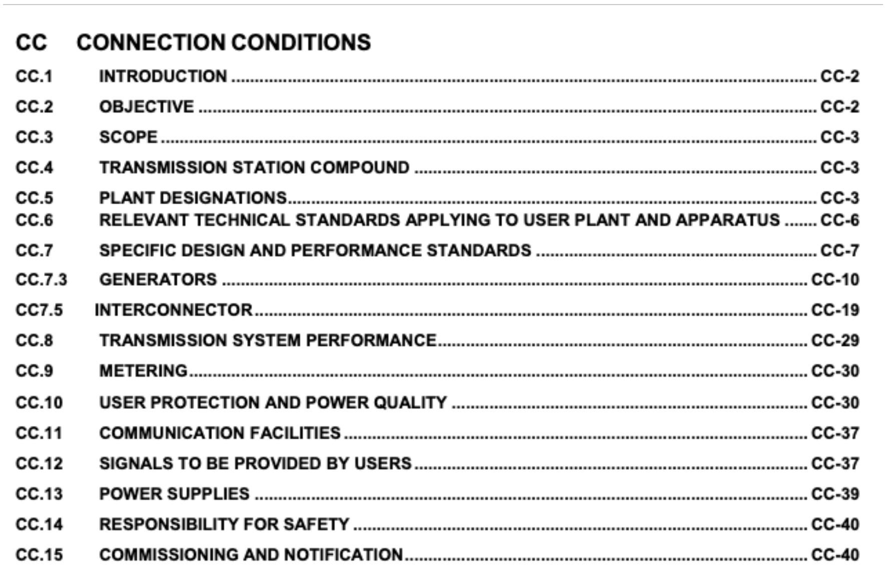
4.2.2 Experiment 1 - Multi-Actor Network Code extraction

In Section 4.1.3, it was demonstrated that a set of distribution codes could be extracted from the document and stored in a digitalised way that could be queried. The relationships between these sets of codes were created and a structural hierarchy was established. For this experiment, the aim is to carry out the same process for the Network Codes so that it becomes possible to align the codes from both sets with sufficient accuracy and coverage. To achieve this, using the same schema defined in 4.1.3, the Grid Codes used by Eirgrid (the Irish TSO) were processed and compared against the ones derived from the Distribution Codes used by ESB networks (the Irish DSO). The aim of this comparison is to ascertain if the hierarchical structure is the same and if a comprehensive mapping between them can be carried out using just the name, without the need for manual examination of the text to confirm the semantic equivalence. For this example, both actor versions of the Connection Conditions section were taken and with the use of queries on the graph database on both graphs demonstrate if a natural mapping can be derived. This choice of example is driven due to their being commonalities in the topics, Metering, Power Quality and Information Required for Connections, that constitute each actor version, another reason that these were chosen is that there is also some actor specific codes contained as is typical of such documents.

The goal of this experiment is to explore the relationships between these codes using visual examination and queries on both sets of concepts stored in the graph database with varying levels of granularity to determine the differences between both sets. This will include determining suitable techniques to extract matching codes from different sets of network codes.

4.2.2.1 Result

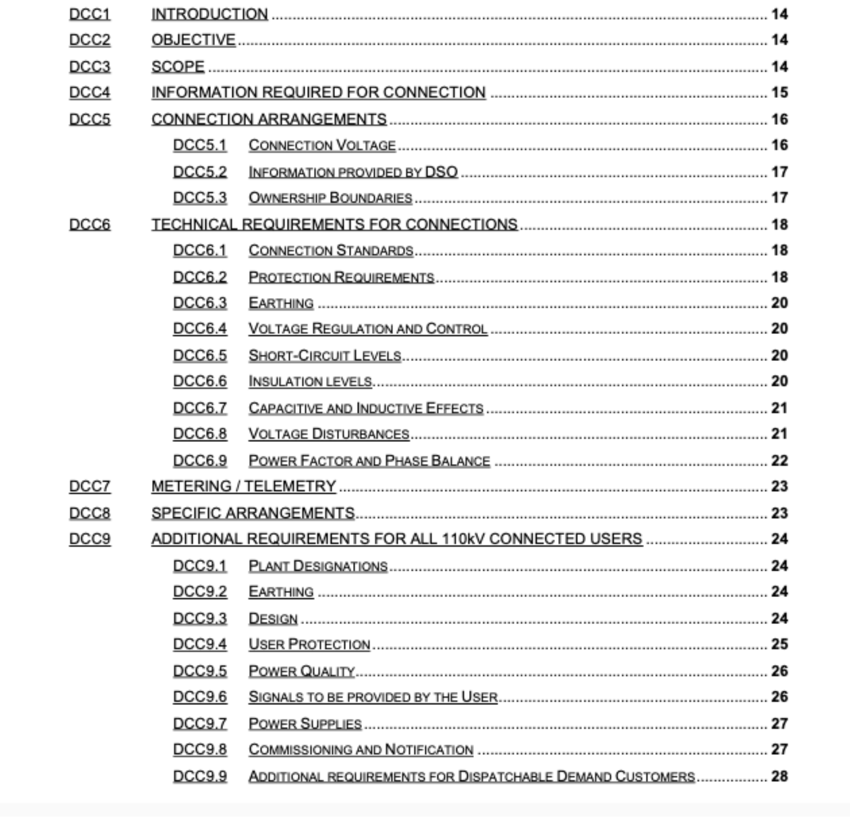
Figures 4.15 and 4.16 are taken from actor-specific versions of the network codes and visually it is evident that there are commonalities between both sets but it is also apparent that there are differences.



CC CONNECTION CONDITIONS

CC.1 INTRODUCTION CC-2
CC.2 OBJECTIVE CC-2
CC.3 SCOPE CC-3
CC.4 TRANSMISSION STATION COMPOUND CC-3
CC.5 PLANT DESIGNATIONS.......... CC-3
CC.6 RELEVANT TECHNICAL STANDARDS APPLYING TO USER PLANT AND APPARATUS CC-6
CC.7 SPECIFIC DESIGN AND PERFORMANCE STANDARDS CC-7
CC.7.3 GENERATORS CC-10
CC7.5 INTERCONNECTOR.......... CC-19
CC.8 TRANSMISSION SYSTEM PERFORMANCE.......... CC-29
CC.9 METERING.......... CC-30
CC.10 USER PROTECTION AND POWER QUALITY CC-30
CC.11 COMMUNICATION FACILITIES CC-37
CC.12 SIGNALS TO BE PROVIDED BY USERS CC-37
CC.13 POWER SUPPLIES CC-39
CC.14 RESPONSIBILITY FOR SAFETY CC-40
CC.15 COMMISSIONING AND NOTIFICATION.......... CC-40

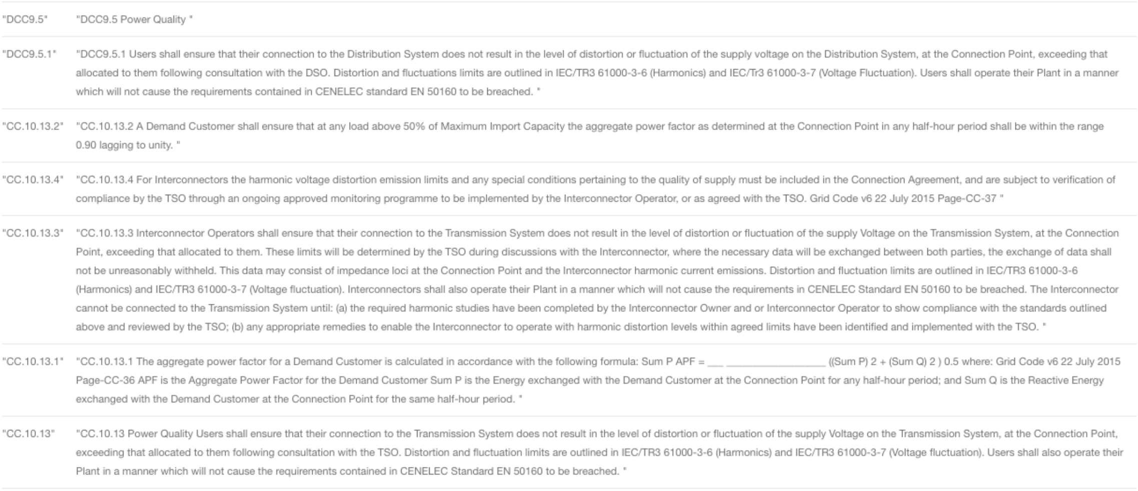
Fig. 4.15 Table of Contents Reference for Eirgrid Connection Conditions



DCC1 INTRODUCTION 14
DCC2 OBJECTIVE 14
DCC3 SCOPE 14
DCC4 INFORMATION REQUIRED FOR CONNECTION 15
DCC5 CONNECTION ARRANGEMENTS 16
DCC5.1 CONNECTION VOLTAGE 16
DCC5.2 INFORMATION PROVIDED BY DSO 17
DCC5.3 OWNERSHIP BOUNDARIES 17
DCC6 TECHNICAL REQUIREMENTS FOR CONNECTIONS 18
DCC6.1 CONNECTION STANDARDS 18
DCC6.2 PROTECTION REQUIREMENTS 18
DCC6.3 EARTHING 20
DCC6.4 VOLTAGE REGULATION AND CONTROL 20
DCC6.5 SHORT-CIRCUIT LEVELS 20
DCC6.6 INSULATION LEVELS 20
DCC6.7 CAPACITIVE AND INDUCTIVE EFFECTS 21
DCC6.8 VOLTAGE DISTURBANCES 21
DCC6.9 POWER FACTOR AND PHASE BALANCE 22
DCC7 METERING / TELEMETRY 23
DCC8 SPECIFIC ARRANGEMENTS 23
DCC9 ADDITIONAL REQUIREMENTS FOR ALL 110kV CONNECTED USERS 24
DCC9.1 PLANT DESIGNATIONS 24
DCC9.2 EARTHING 24
DCC9.3 DESIGN 24
DCC9.4 USER PROTECTION 25
DCC9.5 POWER QUALITY 26
DCC9.6 SIGNALS TO BE PROVIDED BY THE USER 26
DCC9.7 POWER SUPPLIES 27
DCC9.8 COMMISSIONING AND NOTIFICATION 27
DCC9.9 ADDITIONAL REQUIREMENTS FOR DISPATCHABLE DEMAND CUSTOMERS 28

Fig. 4.16 Table of Contents Reference for ESB Networks Connection Conditions

The differences include how the sections containing the codes are structured and also how the codes are labelled. These differences can be overcome by using a simple find query across both graph databases using the labels as query input. The output of the query in 4.17 contains both sets of codes that are relevant to power quality but as can be seen there is a significantly larger body of content in the TSO network codes, those starting with CC.10.13, than in the DSO ones. On further inspection of both grid code documents, it appears that some items relevant to Power Quality contained in CC10.13 and its subsections in the DSO version are contained in DCC.6 and DCC.9.



"DCC9.5"	"DCC9.5 Power Quality "
"DCC9.5.1"	"DCC9.5.1 Users shall ensure that their connection to the Distribution System does not result in the level of distortion or fluctuation of the supply voltage on the Distribution System, at the Connection Point, exceeding that allocated to them following consultation with the DSO. Distortion and fluctuations limits are outlined in IEC/TR3 61000-3-6 (Harmonics) and IEC/Tr3 61000-3-7 (Voltage Fluctuation). Users shall operate their Plant in a manner which will not cause the requirements contained in CENELEC standard EN 50160 to be breached. "
"CC.10.13.2"	"CC.10.13.2 A Demand Customer shall ensure that at any load above 50% of Maximum Import Capacity the aggregate power factor as determined at the Connection Point in any half-hour period shall be within the range 0.90 lagging to unity. "
"CC.10.13.4"	"CC.10.13.4 For Interconnectors the harmonic voltage distortion emission limits and any special conditions pertaining to the quality of supply must be included in the Connection Agreement, and are subject to verification of compliance by the TSO through an ongoing approved monitoring programme to be implemented by the Interconnector Operator, or as agreed with the TSO. Grid Code v6 22 July 2015 Page-CC-37 "
"CC.10.13.3"	"CC.10.13.3 Interconnector Operators shall ensure that their connection to the Transmission System does not result in the level of distortion or fluctuation of the supply Voltage on the Transmission System, at the Connection Point, exceeding that allocated to them. These limits will be determined by the TSO during discussions with the Interconnector, where the necessary data will be exchanged between both parties, the exchange of data shall not be unreasonably withheld. This data may consist of impedance loci at the Connection Point and the Interconnector harmonic current emissions. Distortion and fluctuation limits are outlined in IEC/TR3 61000-3-6 (Harmonics) and IEC/TR3 61000-3-7 (Voltage fluctuation). Interconnectors shall also operate their Plant in a manner which will not cause the requirements in CENELEC Standard EN 50160 to be breached. The Interconnector cannot be connected to the Transmission System until: (a) the required harmonic studies have been completed by the Interconnector Owner and or Interconnector Operator to show compliance with the standards outlined above and reviewed by the TSO; (b) any appropriate remedies to enable the Interconnector to operate with harmonic distortion levels within agreed limits have been identified and implemented with the TSO. "
"CC.10.13.1"	"CC.10.13.1 The aggregate power factor for a Demand Customer is calculated in accordance with the following formula: Sum P APF = ___ ________________ ((Sum P) 2 + (Sum Q) 2) 0.5 where: Grid Code v6 22 July 2015 Page-CC-36 APF is the Aggregate Power Factor for the Demand Customer Sum P is the Energy exchanged with the Demand Customer at the Connection Point for any half-hour period; and Sum Q is the Reactive Energy exchanged with the Demand Customer at the Connection Point for the same half-hour period. "
"CC.10.13"	"CC.10.13 Power Quality Users shall ensure that their connection to the Transmission System does not result in the level of distortion or fluctuation of the supply Voltage on the Transmission System, at the Connection Point, exceeding that allocated to them following consultation with the TSO. Distortion and fluctuation limits are outlined in IEC/TR3 61000-3-6 (Harmonics) and IEC/TR3 61000-3-7 (Voltage fluctuation). Users shall also operate their Plant in a manner which will not cause the requirements contained in CENELEC Standard EN 50160 to be breached. "

Fig. 4.17 Multi Actor Representation of Power Quality Code

In the context of extracting the matching network codes, this single query would not capture all the grid codes related to Power Quality. In order to capture the relevant details across both documents it is necessary to add in some key terms like “lagging”, “harmonic voltage distortion” and “unity”, which are all terms linked to power quality, into the query to expand the result set. These keywords were identified by inspecting the documents and were manually inserted as nodes in the database to provide additional input to queries. Figure 4.18 is the graph that defines suitable actor-specific terms for such items as **earthing**, **design** and **power quality**. While this initial corpus was created manually, in future work this would be driven by NLP and user interaction and validation to provide a more comprehensive corpus of network codes, across multiple document sources.

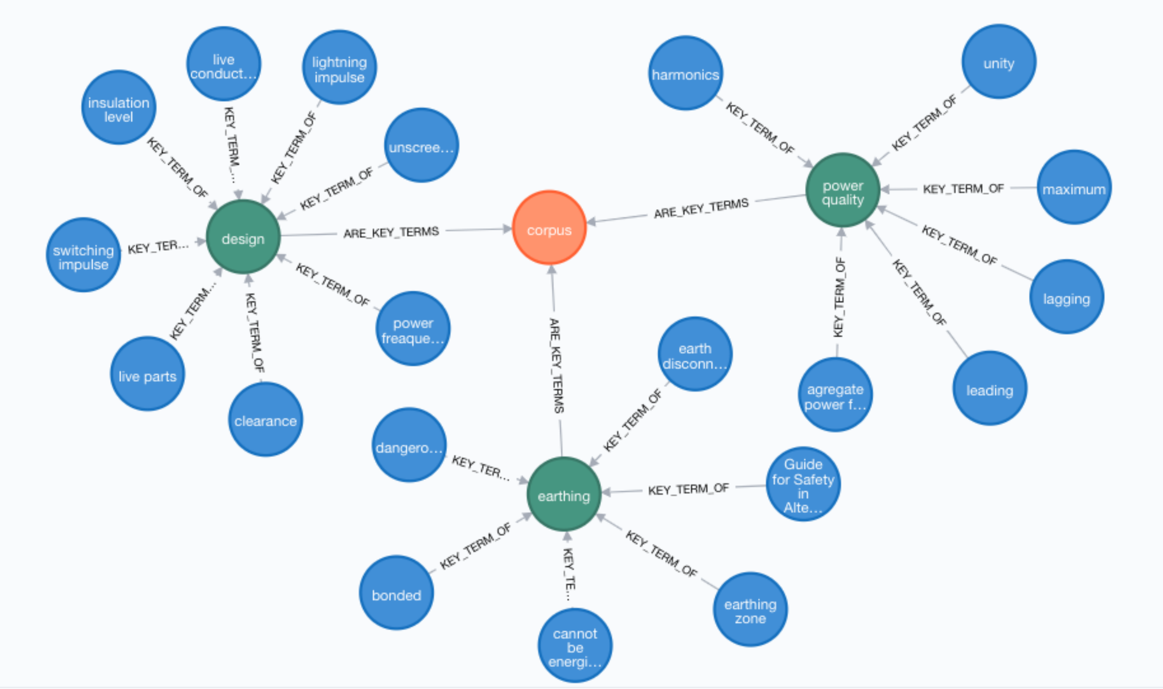


Fig. 4.18 Initial Multi Actor Grid Code Corpus

The result of this query captured the relevant items from the DSO version but also captured some other items from both that were not relevant, for example *DCC10.5.1 Generating Plant Performance Requirements* and *DCC11.4.3 Power Factor WFPSs*, which is centred on Wind Farm constraints, contained in a dedicated section of the TSO version of the codes and has a different meaning in the context of the search. Therefore, querying the network codes using just simple key words is not sufficient to extract all the matching codes based on shared semantics.

4.2.3 Experiment 2 - Precision Concept Mapping Process Definition

In Experiment 1 - Multi-Actor Network Code extraction, we have ascertained that the hierarchical mapping process tested, while high level mapping could be achieved, it is not precise or accurate enough to provide a full alignment across the entire set of codes and it is lack of accuracy that would inhibit the use of this work in a business setting such as the electrical industry which would see its use in critical systems. Experiment 1 - Key Concept Labelling has demonstrated that the codes in a multi actor setting are too complex to perform accurate extractions based on key work searches thus making such a technique not sufficient as a tool to aid the mapping of concepts between multi actor network codes. The following experiment defines and assesses a process that would use Text Similarity and Semantic Similarity as methods to examine the network with the outputs being used and interpreted in a set of business logic that would aim to achieve full semantic alignment of multi actor sets of network codes. While the implementation of Text Similarity and Semantic Similarity are outside the scope of this work the explanation of how its processes and tools can be used to feed the business logic that will define these mappings is not. Business Logic Frankenfield (2019) is a computer program that contains business rules that defines or constrains business operations. In this experiment a set of business rules will be defined using text similarity and sentiment analysis as input with a view towards mapping concepts between multi-actor network codes. To define these mappings it is important to determine what the classifications of these mappings are, should they be presented in mathematical terms, for example if node A is fifty percent the same as Node B, or in more logical terms, for example are two nodes **Not Related**, **Related** or **Equivalent**. Both are valid in terms of classification that could be applied but for the purpose of clarity and clearly defined clustering in this case a logical mapping is chosen. Figure 4.19 details the proposed flow of how the concepts, once extracted, can be mapped across multiple actor versions using relationships to define Semantic and Text similarity and using the properties of those relationships to classify the mapping.

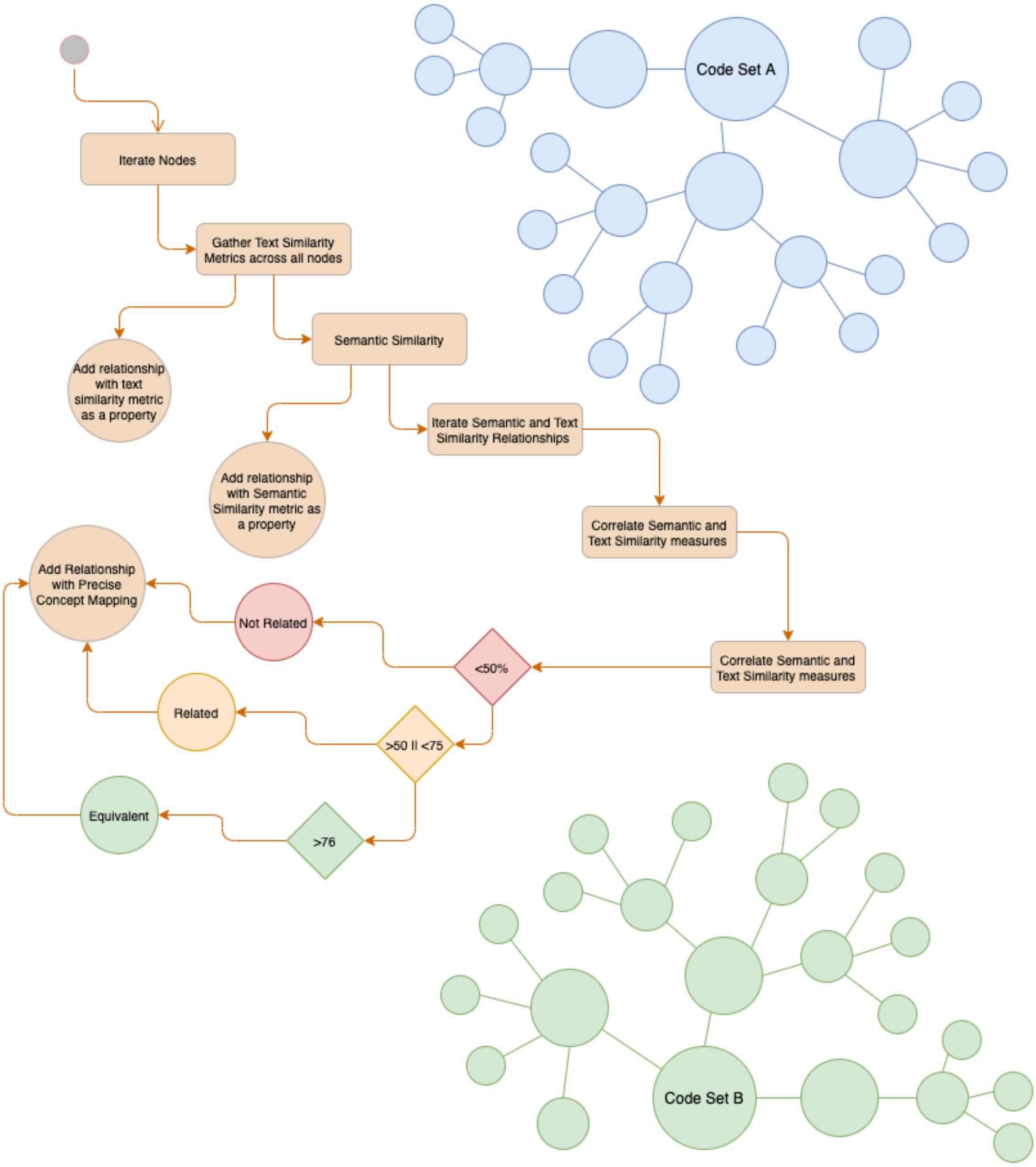

Iterate Nodes
Gather Text Similarity Metrics across all nodes
Add relationship with text similarity metric as a property
Semantic Similarity
Add relationship with Semantic Similarity metric as a property
Iterate Semantic and Text Similarity Relationships
Correlate Semantic and Text Similarity measures
Correlate Semantic and Text Similarity measures
<50%
Not Related
>50 || <75
Related
>76
Equivalent
Add Relationship with Precise Concept Mapping
Code Set A
Code Set B

Fig. 4.19 Precision Concept Mapping Classification Flow

This flow involves iterating across all nodes and comparing the text for similarity with all nodes in the other set with the result being added as a property of a relationship called SIMILARITY. The next step involves, again iterating over all nodes and this time, comparing the nodes for Semantic Similarity with the scores again being added as a property of the SIMILARITY relationship. It is this SIMILARITY relationship and its properties that

Table 4.1 Grid Code Text Analysis

Experiment	Text 1	Text 2	Word2Vector	SequenceMatching	Correlated (%)
1	dcode	dcode	1	1	100
2	dcode	gcode	0.9644547	0.0090514	87.296
3	dcode	notcode	0.9448368	0.0075578	71.408
4	dcode	notelectrical	0.7203368	0.0020410	14.695

Table 4.2 Grid Code Text Clustering

Experiment	Text 1	Text 2	Correlated (%)	Cluster Group
1	dcode	dcode	100	Equivalent
2	dcode	gcode	87.296	Equivalent
3	dcode	notcode	71.408	Related
4	dcode	notelectrical	14.695	Not Related

form the input for the classification algorithm and determine the nature of the mapping between both sets of concepts. The next step involves correlating both similarity scores and converting them to a percentage which is then used to classify the mapping between the two connected nodes.

Taking the example of the features extracted that are related related to Power Quality we used in 4.2.2 and analysing the text in terms of Text and Semantic similarity using Sequence Matching and Word2Vector (Mikolov et al., 2013) respectively we conducted the following experiment.

We analysed four different samples for this, samples exactly the same text as a control, samples with similar concepts, samples from the same genre (electrical engineering) but should have minimal equivalence and totally non related text. For the purpose of this experiment we have labelled the samples as *dcode*, *gcode*, *notcode* and *notelectrical*.

The results of these experiments are presented in table 4.1 and, as expected, the comparison between dcode yields a result of 1 on both counts which indicates that both processes are working. In interpreting the results of experiments 2, 3 and 4 it is evident that the content can be compared for similarity with experiment 2 containing content that matches in terms of meaning and word similarity as it scored highly in comparison to the other two samples. It is also evident that electrically themed wording can be distinguished from non electrical themed text upon comparing the scores between experiments 3 and 4.

Based on the metrics above the text analysed would be clustered in the following way.

4.2.3.1 Result

While the results presented in tables 4.1 and 4.2 have been gathered by performing analysis on a small subset of the network codes it is evident that there is a correlation between text similarity and semantic similarity that can be used for concept mapping purposes. While conscious of the very small subset and the basic similarity analysis carried out, the reason for this analysis is to provide business logic to the process that assigns relationships. As expected, the non electrically themed text would have a very low similarity score of 14% which would justify the mapping to the *non related* cluster while the text extracted from the grid code document but from an unrelated section contained enough similar terms and phrases to justify it getting assigned to the *related* cluster. What is telling is that even though the Text 2 sample in experiment 4 was a combination of multiple codes which constituted the code in Text 1 the similarity was striking with 87% similarity detected which justifies the assignment to the *equivalent* cluster.

Given these results it is shown that the proposed system can perform an initial clustering of the concepts but if full mapping were to be achieved across the document the initial clustering would need to be checked for accuracy by a human expert and the interpretation of the similarity metrics and the business logic might require re-factoring. While agreement between text similarity and semantic similarity was not achieved in all cases across both documents, it should be noted that there is flexibility to tweak the similarity metrics, their thresholds and the business logic in a way that improves their agreement with the semantic alignment between concepts identified by a human expert. The algorithms for making this improvement are assessed in 4.2.4 and combined with business logic that offer a nuanced understanding of how two codes can agree, offers a way to achieve scalable automated semantic alignment between network codes from different stakeholders.

4.2.4 Experiment 3 - Assess the matching of concepts from multi-actor concepts based on distance and semantic similarity

In the previous two experiments we have defined the relationships required and the processes needed to form digitalised sets of network codes upon which deeper analysis can be carried out and new relationships can be discovered between multi actor sets of network codes. While full semantic alignment cannot be achieved at this stage some of the techniques that may be used to achieve this can be assessed. In Scenario 2 we presented a fictitious situation where a change in how codes relating to data are presented in the text, the sample codes used to analyse these techniques are based on the presence for the word "data" in the network code text. The 4 techniques we use in the experiments are Jaro Winkler (Black, 2021) Distance, Sorenson Dice Similarity (Neo4J, 2021), and the Bag of Words Based Semantic Similarity and Word2Vector models and tools provided by the Gensim Project (Řehůřek and Sojka, 2010). The list of techniques chosen are non-exhaustive and are just a sample of those available and there are other additions, like Term Frequency-Inverse Document Frequency (TF-IDF) by Scikit-LearnPedregosa et al. (2011) and Natural Language Processing (NLP) for example, that could also be investigated, this would be a more focused analysis in tandem with the definition and investigation of a corpus to both feed and be fed by the analysis. To begin the experiment we created a relationship, **HAS_DATA_IN_DESCRIPTION** between all the concepts from both the distribution and grid code groupings that contained the word "data" which resulted in about 4000 relationships. This means that if the word data appears in the code in set A (distribution code version) it is related to every other code in set B (grid code version) but it is necessary to make this relationship so that there is no loss in terms of the potential links that could be made in the assessment of each of the four techniques. To verify this a visual search was carried out across both sets for codes that contain the word "data" and as seen in figure 4.20.

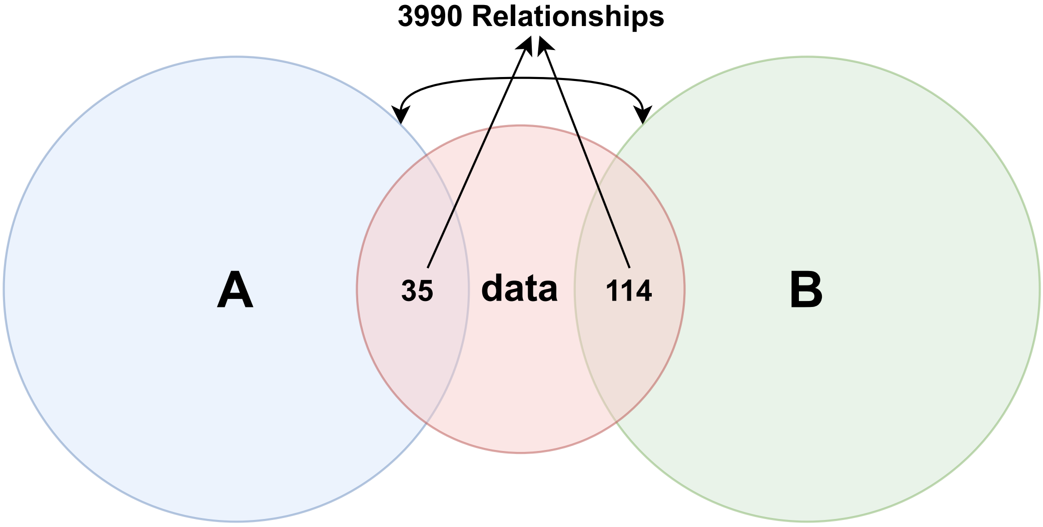


Fig. 4.20 Breakdown of relationships between sets

The next step was to run each techniques with the sample taken predicated on the concepts being related through the **HAS_DATA_IN_DESCRIPTION** relationship. The outputs of the techniques are added to the relationships then to define the strength of the relationship between one concept and the other based on the technique used. The key element of this experiment is to compare the outputs of the techniques applied for effectiveness and precision in how they accurately they identify the strength of the relationship between one concept and another. This comparison was carried out by computing the scores and ordering them with a view to finding a cut off point where the scores were weak enough for them not to show any relevant similarity and then visually inspecting the text for the highest scoring relationships to determine the best performing technique or combination of techniques. The plot in figure 4.21 contains the scores for all four techniques ordered descending from the most similar to the least similar. Firstly, it must be noted that all traces have a different scale of deviation from the most similar, with Semantic Similarity reamining high and Sorensen Dice dropping significantly, and also the steep drop off in all scores which would indicate that there are some relationships very strong and the rest not so much.

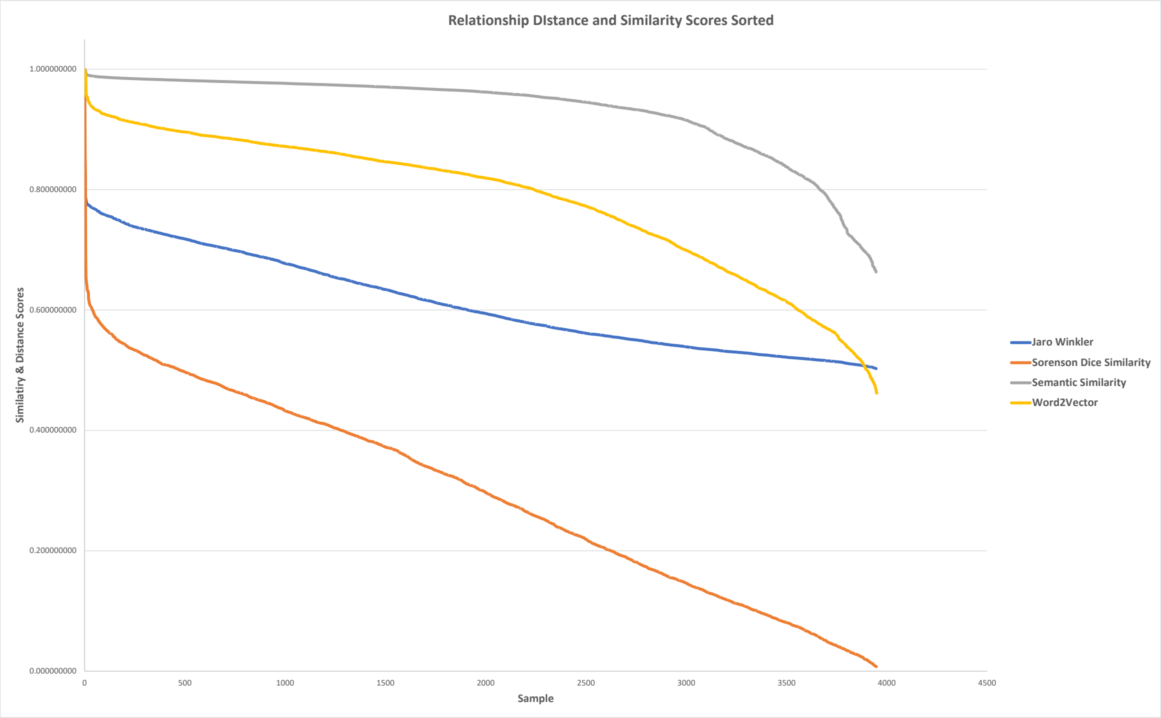


Fig. 4.21 All Similarity and Distance scores ordered

To explore this we will look at each trace in isolation and take a smaller sample of the twenty most similar relationships.

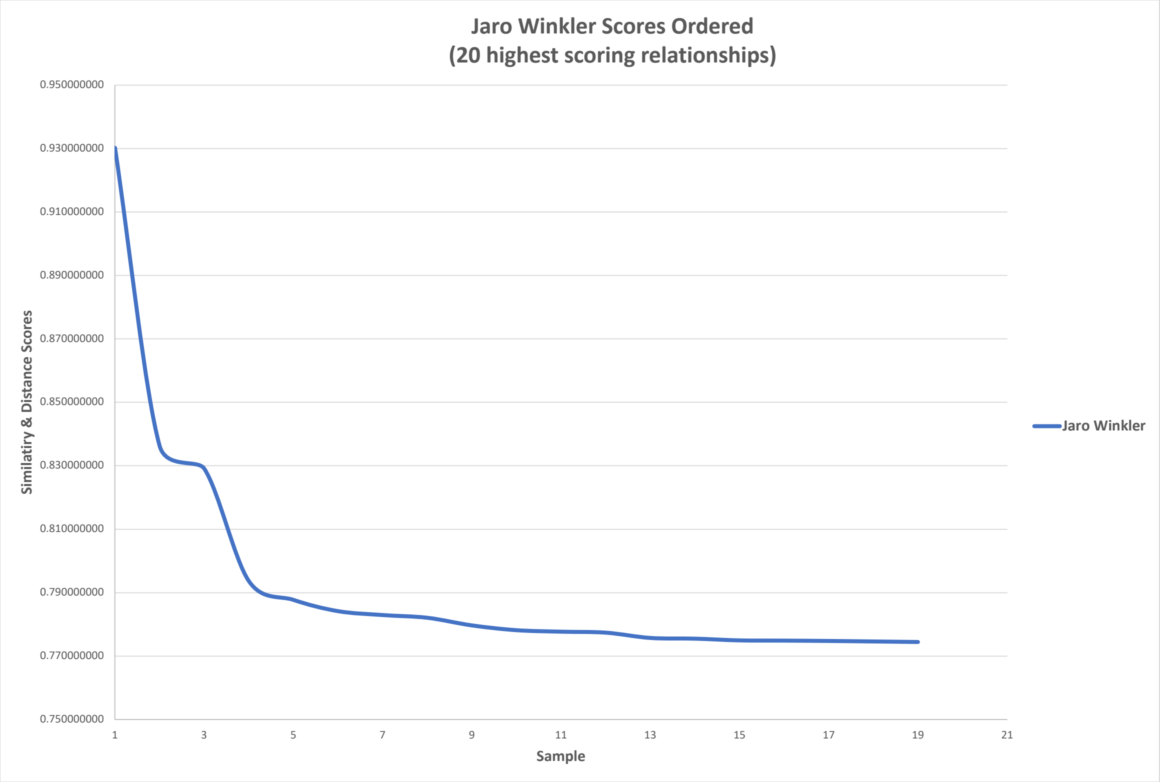


Fig. 4.22 Jaro Winkler Distance scores ordered

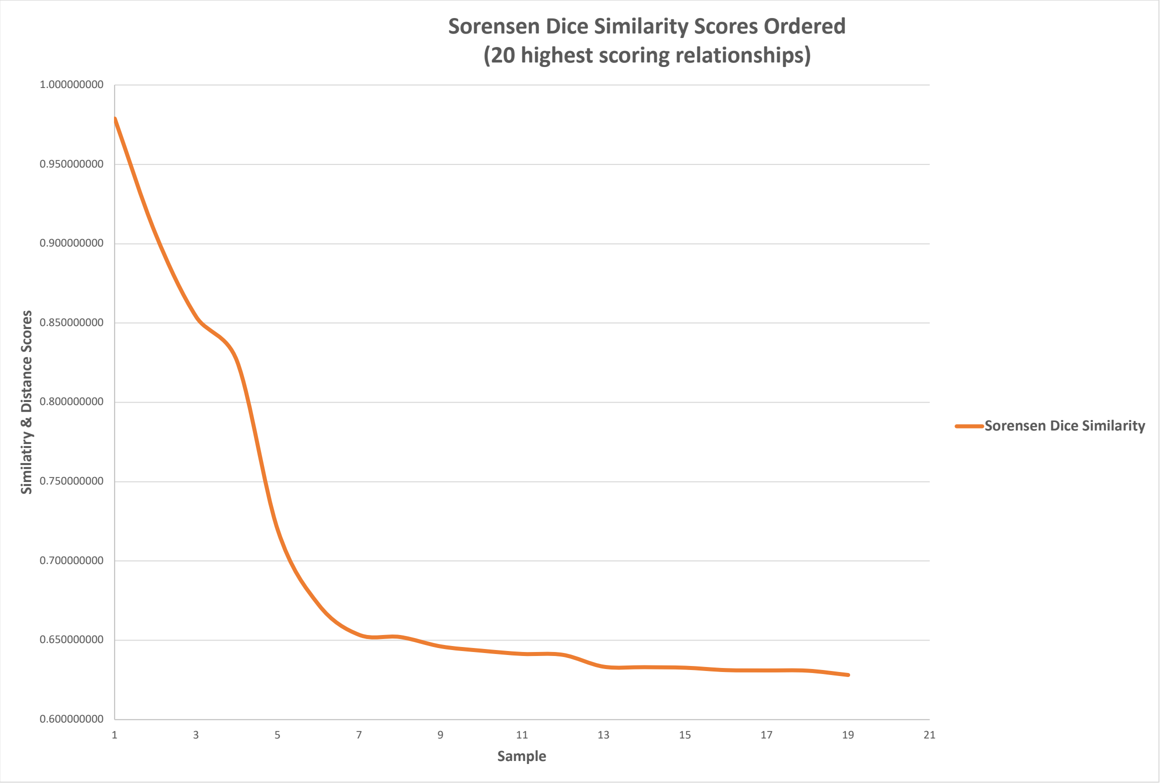


Fig. 4.23 Sorensen Dice Similarity scores ordered

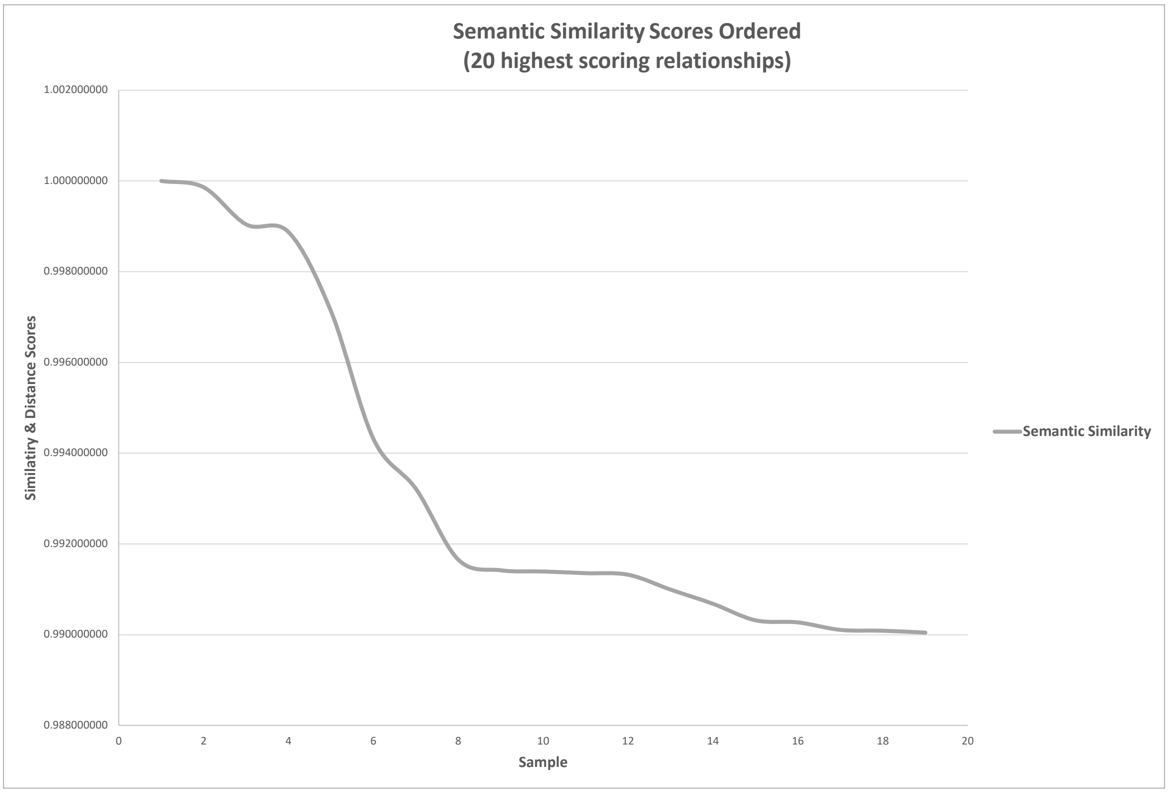


Fig. 4.24 Semantic Similarity scores ordered

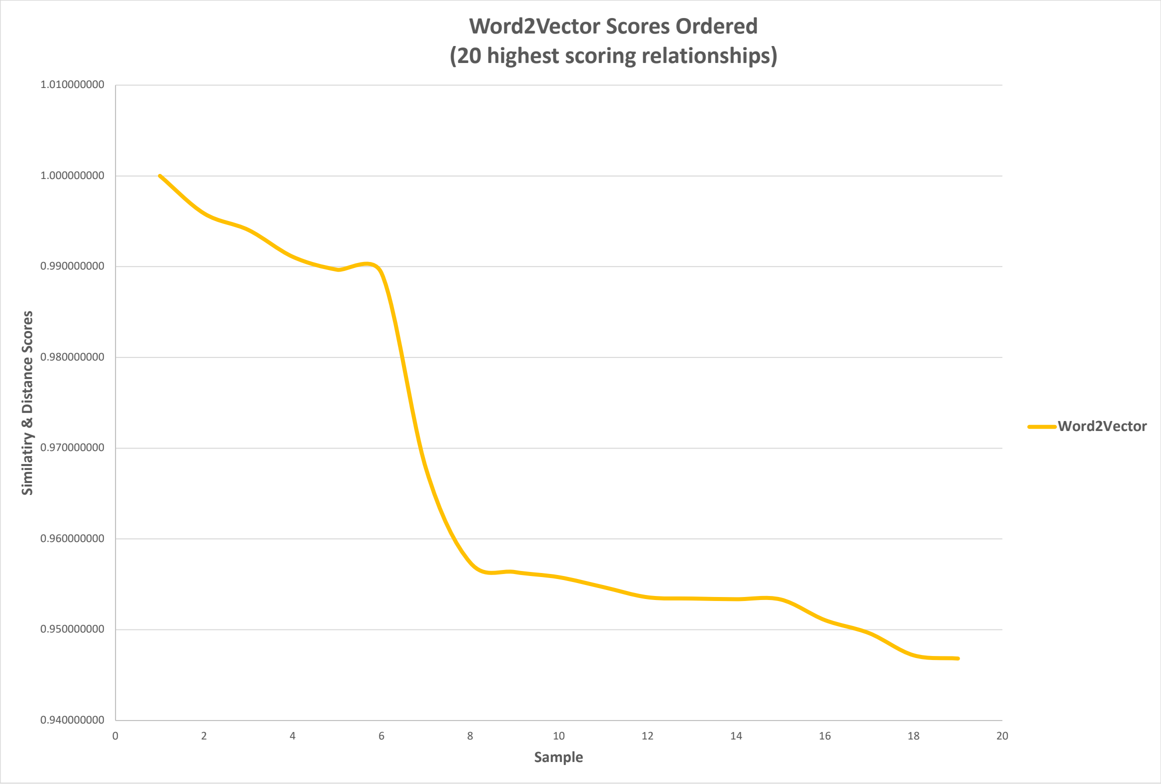


Fig. 4.25 Word2Vector scores ordered

The first thing to note is that in both the Jaro Winkler Distance scores ordered and the Sorensen Dice Similarity scores ordered plots begin below 1.0 and decrease quickly where as the plots for Semantic Similarity scores ordered and Word2Vector scores ordered begin at 1.0 and do not decrease as quickly. This is due to the presence of two concepts sharing an equivalent term "dcc9.9.3 supervisory control and data acquisition (scada)" in set A and " oc7.2.5.3 supervisory control and data acquisition (scada)" in set B. Because both Jaro Winkler and Sorensen Dice look at words and not the patterns of words and determined that while similar they were not equivalent based on the differing code numbers being present in the text which is strictly correct if the goal was to exactly match the text word for word. Both Semantic Similarity and Word2Vector scored this as 1.0 due to the context and the word pattern of both pieces of text being equivalent. In viewing the shape of the plots it is also noted that there is an intermediate elbow and a knee in, some more pronounced that others, that occur between numbers 3 and 5 in the sample. This is less pronounced with the plots for semantic similarity and word2vector that it is in the other two and does not result in the score decreasing to the same degree. When visually inspecting the 4th relationship which is between to bodies of text, one for code DCC11.6.1 Wind Turbine Generator Dynamic

Models and the other for code PC.A4.10.1.1 MODELLING REQUIREMENTS FOR WIND TURBINE GENERATORS, it is noticeable that they contain common words and even phrases but the text differs and the length of the passage differs. They essentially govern the same thing but code PC.A4.10.1.1, which is belonging to the TSO has more specific to the connection of wind driven generation assets to their system. With the scores for Jarow Winkler:0.7070731901937156 and Sorensen Dice:0.7196467991169978 marking lower than Semantic Similarity:0.9988691210746765 and Word2Vector:0.9896563 it would lead again to the purely text based algorithms basing their scoring on word count and text length where as the Vector and Model based techniques are more subtle in identifying phrases and word groupings as similarities.

DCC11.6.1: ***wind turbine generator*** *dynamic models the tso requires suitable and accurate dynamic models for all generators connected to, or applying for a connection to, the transmission system or the distribution system* ***in order to assess reliably the impact of the generator's proposed installation on the dynamic performance and security and stability of the power system****. modelling requirements for thermal and hydro generators are processed on the identification by the applicant of the relevant pss/e 9* ***library model and the provision of the applicable data parameters*** *as set out in dcc11.3.4. where there are no suitable library models available, specially written models are supplied. these are known as* ***"user-written models"***

PC.A4.10.1.1: the tso requires suitable and accurate dynamic models for all generators connected to, or applying for a connection to, the transmission system ***in order to assess reliably the impact of the generator's proposed installation on the dynamic performance and security and stability of the power system****. modelling requirements for thermal and hydro generators are processed on the identification by the applicant of the relevant pss/e* ***library model and the provision of the applicable data parameters*** *in the current, appropriate application form. where there are no suitable library models available, specially written models are supplied. these are known in pss/e as* ***"user-written models"****. currently (september 2004) there are no suitable pss/e library models for* ***wind turbine generator****s. as a result, the tso requires controllable wfpss greater than 5mw to provide specially written models and associated data parameters specific to the* ***wind turbine generator****s and any associated controls and reactive compensation equipment to be used in the applicant's controllable wfps scheme. the requirements of these models are as outlined in this section of the planning code appendix.*

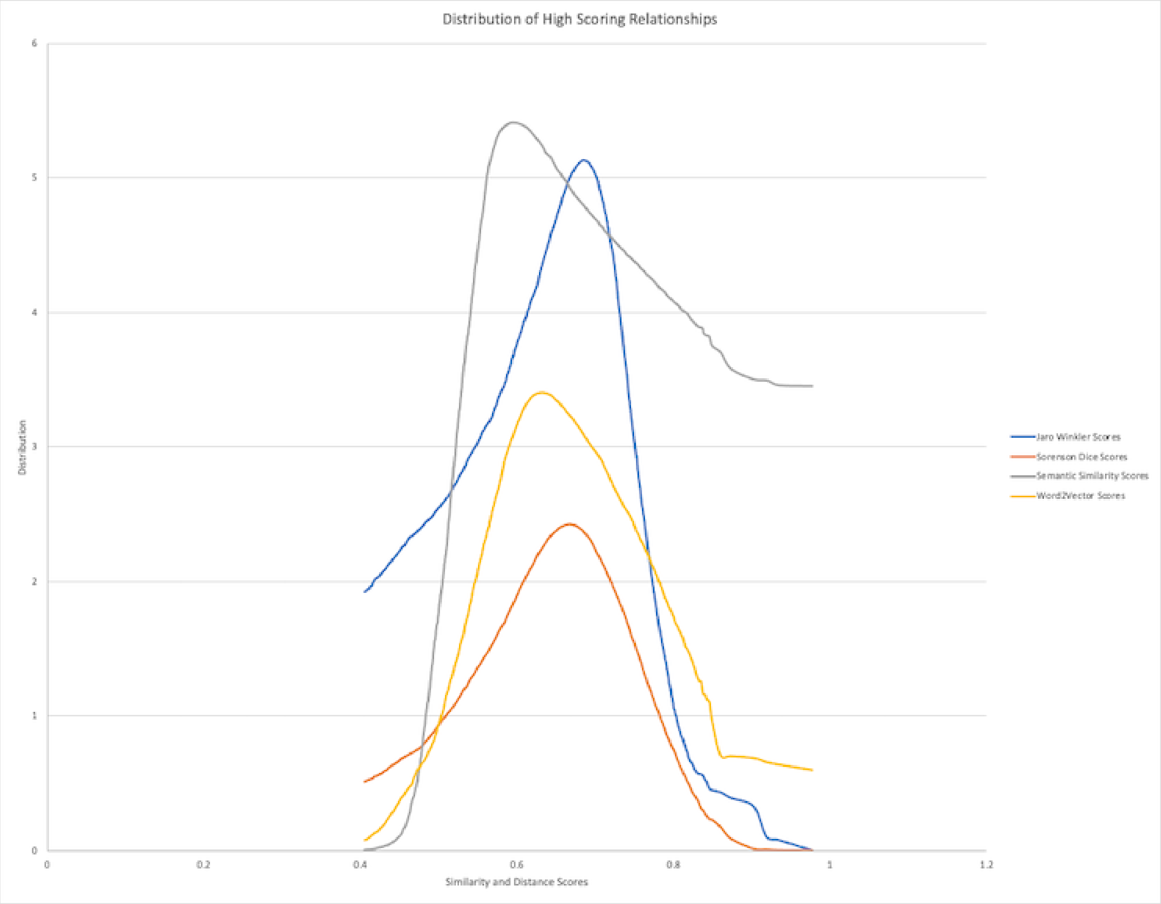


Fig. 4.26 Distribution of High Scoring Relationships

4.2.4.1 Result

The results presented above prove that all four techniques can match the concepts with Jaro Winkler and Sorensen Dice being less effective due to their lack of phrase analysis. However, in viewing the trace for Jaro Winkler in figure 4.21 it could be used to perform a preprocessing step by identifying the most likely candidates for deeper analysis as it remains relatively flat across the entire dataset compared to Sorensen Dice. This is seen in the plot in figure 4.26 that looks at the distribution of scores across the techniques where the peak of the distribution of the Jaro Winkler scores is at the upper end of the scores and the tail of the graph towards the lower scores is still quiet high even considering we discarded any relationships that had less than 0.4 in their scores.

Based on the scores and visually inspecting the matched codes identified either of the algorithms would be suitable for use in a part of the process in the Experiment 2 - Precision Concept Mapping Process Definition experiment but to say that it would definitively match all codes with a degree of accuracy and precision would be incorrect. The expert would still need to visually verify the matchings as a system and a non expert might draw false assumptions and this is evident when visually inspecting the matched codes, particularly the distribution codes, where there is a considerable amount of codes being referred to by other codes that may lead to the assumption that the content in a matched code may spread across more than one code. But this is not a dependable assumption as in the case of the two sample codes above where dcc11.3.4 is referenced in DCC11.6.1 and relates to Ramp Rates, which have no obvious linkage in terms of wording but maybe the link is implied by an expert's interpretation.

The conclusion of this experiment is that these algorithms can match codes with varying degrees of success but they cannot perform a comprehensive precise and accurate matching across the entire set of codes and to perform this the expert would need to verify the results and feed those results back into the algorithms and the queries running them in terms of thresholds and potentially referenced codes that need to be joined to form the full code that has to be matched. This would be for the expert to decide and, as is common practice in software architecture and engineering where the knowledge of the end user and their requirements is needed to develop systems, a set of user centred rules would be developed to aid the user, in this case the expert, define a process that can be translated into software to verify and make recommendations in a future code matching system. These rules could be developed in JBoss Rules, SPARQL or as a set of Behaviour Driven Development scenarios developed in Gherkin like so.

```
Feature: Expert wants to match codes across two sets of documents
  In order to make a coherent and accurate set of multi actor codes
  As a Grid Code Expert
  I want to gain match codes across two sets of stores concepts

  Scenario: Expert visually inspects similar codes to verify accuracy
    Given there are two sets of codes as nodes in a graph database
      And a relationship has been created
       between nodes based on the presence of a term
      And a set of similarity and distance scores
       have been determined and added to the relationship
     When the expert views the codes with the
       highest similarity and distance
    Then the expert should be able to verify the
      accuracy of the similarity score
      And remove the relationship if the
        similarity score provides a false match

Scenario: Expert wants to make a match between a
           combination of codes referenced
    Given there are two sets of codes as nodes in a graph database
     And a relationship has been created
      between nodes based on the presence of a term
     And a set of similarity and distance
      scores have been determined and added to the relationship
    When the expert views the codes with the
      highest similarity and distance
      And the expert identifies an referenced code
    Then the expert should be able to query the referenced code
      And add a relationship if the reference code
        is part of the match

Scenario: Expert wants their verification
           and new relationships captured
   Given there are two sets of codes as nodes in a graph database
     And the expert has verified the results
     And the expert has created the relevant
      relationships based on identified referenced codes
    When the expert re runs the original similarity matching
    Then the new relationships and the actions taken when
      verifying the similarity based matches
      should be inputted into the algorithms and queries
```

The rules defined are not a concrete set of rules but rather a sample of what may be developed and if perused a concrete set would need to be developed in conjunction with the expert(s). Any such systems built from these scenarios and rules would form the interface where by the expert input would combine with the similarity and distance analysis and the process defined in 4.2.3 to form a system that would effectively match the codes from both documents.

Chapter 5

Conclusions

5.1 Summary

In both scenarios we looked at the practical application of a novel representation of the network codes and while these scenarios framed the experiments and went towards answering the research questions it is important to view the results and the work in terms of the challenge and the contribution of the work furthering the state of the art. The challenge came from large scale developments of the electrical supply industry: energy is produced, consumed and monitored in a more distributed fashion and digitalisation is helping power systems to adapt to these changes. This was further highlighted in the background section (Section 1.2.2) which identified the current process of updating and managing the network codes as being a manual one that is carried out on static documents, requiring expert knowledge. Compare these manual operations and the current textual representation of the network codes with the current state of the art identified in the literature review of Section 2.2 and it is clear that the network codes are falling behind and have the potential to inhibit potential benefits like the use of autonomic grid management, which might otherwise provide an answer to the growing complexity of the electricity supply system.

In Chapter 2 we explored the state of the art in Section The Digitalisation of Traditional Documents, how knowledge is represented in the smart grid (Section 2.2) and the use of graph similarity (Section 2.3) for the purpose of understanding the state of the art and of identifying gaps, some of which are addressed in this dissertation. We noted that there is a large body of research into the use of graph similarity, matching and recommender systems in applications like computer vision, chemistry, social networks and biomedical systems, but their use is relatively novel in smart grids. Chapter 4 describes experiments, for which the main goal was to address the research questions in the context of the scenarios. However, the experiments also investigate processes, building on the state of the art in Section 2.3, to advance the state of digitalisation and knowledge representation in the smart grid.

Section 3.1 states that, to overcome the gap between the potential of digitalisation and its current implementation, *a set of technological processes using text mining, semantic alignment and enhanced digital representation of the Network Codes will help provide a more semantically aware, digitised and interactive version of the network codes* and by posing research questions (Section 3.2) and answering them through experiments (Chapter 4), the process we investigated is that of digitising traditional documents by extracting concepts from them and representing them in a way where the content can be queried and analysed in a more dynamic and deeper way than the traditional document based representation. By so doing, we developed a new representation of knowledge in the electrical supply industry and by exploring a new use case for using graph similarity in the representation of digital information. Also in Chapter 4, the outcomes of the experiments were assessed and it was found that by using text mining, basic semantic and text similarity, and exploiting the features offered by graph database storage that it was possible to create a novel representation of the network codes that could be used in a broader range of scenarios than the text based versions. This is due to the representation created being more machine readable due to being queryable in a way that is machine driven thus allowing user interfaces and software components to be built on top of the representation and also due to the data being stored in a way that can facilitate a deep analysis of the data that can facilitate, among others, ways to define relationships between codes even if the codes are in multiple documents.

In assessing the overall outcomes of the experiments it is important to view them in terms of

- are they
  - usable
  - repeatable
  - reliable
- and can they be built upon to further the state of the art.

Usability in this case centres on how usable are the new representations of the codes to the computer systems that may want analyse and query them and based on that the novel representation of the network codes are at least as usable as their predecessors but with the added functionality of being able to perform adhoc querying and to enable new and diverse relationships to be developed between other network codes both within the same organisation and across representations used by other Actors in the grid. These results are also repeatable, as defined by Plesser (2018), in terms of a formal software process being developed and a repeatable deployment mechanism created to enable the network code mining process to be deployed in multiple environments.

Regarding reliability, the experiments extract the concepts and from the experiment assessing the impact of change, where an annex item was represented in a number of different ways it is clear that further work is required to make semantic alignment work better because some alignments could be missed, but it must be noted that the process framework can be extended with new business logic and the ultimate goal would be to implement natural language processing (NLP) more fully.

The degrees to which the experiments answer the research questions vary and two common limitations were evident, the lack of a corpus and the need for the expert to help the system. RQ1, *How do we represent the data from a Network Code document so that a subset can be used?* and its sub questions are mainly centred on single actor codes and the extraction of the codes from the documents and investigating a method to store them so that more advanced operations, like assessing the impact of change, can be carried out.

To address these an experiment carried out to answer each sub questions in section 3.3.1 which has the overarching goal of assessing the impact of change specifically in the Distribution Codes. In Experiment 1 - Key Concept Labelling the aim of the experiment was to extract the sections and sub sections of the document and label them. This was proven to be successful and the process illustrated in figure 4.6 was used to successfully extract and label the concepts from the document with a degree of accuracy. While the process worked it was found that the initial stage of the concept extraction required some mappings that were fed into the system to identify the high level sections. Experiment 2 - Creating A Schema centred on the creation of a schema, and exploration of a database technology best suited to host such a schema. A core aspect of this experiment was to ensure that the concepts were stored in a way that the text and their relationships with the other codes were maintained. It was decided that the functions of Graph Database technology would be most suitable as it provides a strong focus on the relationships between concepts and also how the nature of the relationship can be analysed. Each concept was stored in a node with the text as a property of the node and based on the position of the concept in the document a relationship was created based on if it was catagorised as or a subsection of another node. This schema was found suitable and in subsequent experiments was used as the basis for the remaining analysis. In Experiment 3 - Assessing the impact of change, using the stored concepts the from the previous two experiments we attempted to assess the impact of a change to an annex item that was referenced throughout the document. We first used a basic "find" function from a pdf viewer to find the references as a control and then used database queries over the stored concepts to find the references there. This was successful and through this experiment it was noted that the annex item was referenced in a number of different ways through out the text. This enabled the enhancement of the query to account for the references which is a function not readily available in pdf viewers. This experiment was carried out on a relatively small dataset but across a larger dataset with a more complex assessment of the change the enhanced querying available would lead to a faster and more comprehensive assessment of the change. To summarise, *RQ1: How do we represent the data from a Network Code document so that a subset can be used?* has been answered because it was possible to prove that a representation of the Network Codes could be created that can be labeled, stored and assessed as either an entire set or a subset.

In section 4.2 the aim was to attempt to answer RQ2, *Given there may be more than one set of Network Codes, how do we find commonalities across them?* and its sub questions and this was framed in a scenario that looked to find concepts across both sets of codes. To address the practicalities of the scenario there were three experiments carried out that explore the extraction of the codes from more than one set of documents, define a process to map them and assess some techniques that can find similarities between them. In Experiment 1 - Multi-Actor Network Code extraction a set of concepts were created from both sets of documents and the aim of the experiment is to ascertain if the concepts were extracted accurately and it was possible to query across both sets of concepts for codes with a common term, "Power Factor". This was successful but it was identified that a simple query on a search term was not sufficient to capture all codes related to the common term as some of the concepts were missed as they refer to some of the terms related to Power Factor like “lagging”, “harmonic voltage distortion” and “unity”. For this an initial corpus was created and, as identified, this corpus would need to be expanded and used in conjunction with similarity algorithms and potentially NLP to form a fully functional system. Experiment 2 - Precision Concept Mapping Process Definition centred on devising a process that would enable the mapping of the codes across both sets of concepts. This experiment was to outline the steps and propose three relationships that could be applied to the codes to signify the degree of similarity between one concept and another. Here the first introduction was made to using similarity and distance algorithms to help identify similar codes. It must be noted, that while the process was identified the thresholds upon which the relationships were defined warrant deeper scrutiny and in Experiment 3 - Assess the matching of concepts from multi-actor concepts based on distance and semantic similarity a set of techniques that identify similarity and distance between two text elements were examined on a common set of concepts. While all four techniques were potentially suitable with varying degrees of success it was identified that the expert would still be required to identify commonalities between codes but the process defined and techniques examined would be capable of providing a starting point for the expert.

The schema, derived relationships and the features captured by the graph model can already be used for impact assessment within a network code document, and to extract related codes from multiple documents, as was shown in scenarios 1 and 2 respectively. To serve as a basis for further work however, a deeper level of validation would be required from the electrical network planners to assess whether the outputs, as they stand, are sufficiently reliable in the sense of being both complete and correct. This would require extensive validation and co-development of the framework and goes beyond the scope of this dissertation.

5.2 Future Work

The previous section claimed we had made some progress beyond the state of the art in terms of digitising the network codes but that robust validation, by domain experts refining the mapping between concepts, would be needed to go further with this line of research. The starting point for this would be the development of a query interface that would allow the network planners and power system engineers to issue a query based on a search term and in return receive the relevant concepts. The concepts returned would then be validated in terms of suitability and the results used to augment the initial corpus developed in the experiment in Section 4.1.3 and then used to weight the relationships developed in the experiment (Section 4.2.3). This would involve extending the business process built on the iterative approach based on graph similarity and to incorporate user-based collaborative filtering (Harpale and Yang, 2008) This would provide an autonomous method of validation for the network codes and allow supplementary documents, like standards, to be ingested, parsed and used with related network code concepts.

One method of validation would be through the deployment of the system as an open platform or the dissemination of the system built to run the experiments through the open source community. This would involve providing a mechanism where the experts and scientific community alike could reproduce the experiments and allow them to review and validate the representation of the codes, reproduce the experiments with a larger sample of data and also allow the research and scientific community use the platform to perform a more involved experiments matching novel use cases.

When validated, the new network code representation has the potential to enable more engineers (and not just network planning experts) to consult the network codes and to use them in an informed way. For example a scenario like the connection of a significant DER cluster could be presented in terms of the potential network code violations that could occur by a system built to query the new representation by the key words contained in the scenario, providing a power systems engineer with the ability to make informed choices based on information presented in an accessible way other than having to research or engage with an expert in network codes to determine which codes are relevant to the use case.

A further use case for a fully expert validated set would be centred on the extraction of quantifiable values that could be used in system monitoring and control techniques driven by a Policy Based Grid Management based system that ingests the values and uses them to monitor grid control techniques as proposed by Ryan et al. (2019).

Another such scenario would be centred on the changing complexion of the energy sector and the devolution of responsibility from the traditional DSO/TSO model to a more distributed model where the expertise may not be present in the management structure, for example, a Virtual Power Plant (VPP). To compensate for the potential lack of domain expertise here the codes specific to the connection of VPPs to the network could be presented based on a role specific to VPP owners and by doing so lessen the reliance on needing an employee or a consultant with an in depth knowledge of the network codes to determine the regulations specific to the running of the VPP from a regulatory perspective.

A further step might be to investigate how to support “what if” analysis in the model. This would involve creating and implementing a new use case based on the new representation and introduce a new standard or regulation that might impact more than one code. The realisation of this use case would provide the network codes with a level of flexibility and add an extra layer of control for the electrical engineers who would integrate such regulations and standards and by doing this enable the network codes keep pace with a rapidly changing industry.

Further research in this area could include investigating the potential of extending the system to involve such functions as conflict identification, resolution and the highlighting of non-binding constraints that would result in a more rigid, robust and resilient set of network codes. Furthermore such a research could form a digital twin of the network codes that may form part of a regulatory sand box where new regulations could be trialed, proposed and their impact assessed with a layer of separation created between this representation and the actual validated and agreed upon codes.

The two previous future work items (devising a robust validation strategy, and supporting "what if" analysis) relate to the network codes themselves but there are also open questions around how to manage multiple network code documents and the standards they reference. Section 4.2.3 outlines a new process that requires stronger validation than is currently possible but has enormous potential for business flexibility. Given suitable concept mappings and term substitutions, a robust federated model could be devised, in a form that could be shared and managed by the relevant stakeholders. Ultimately, this could be extended to broad geographical regions, perhaps even the EU itself, with benefits to smart grid stability and cross-border energy trading. The research challenges to support this would focus on improving the scalability (in terms of numbers of codes and specification formats) and on identifying ways of collecting the relevant information from the stakeholders and weaning them off their reliance on static documents with limited support for interoperability, and onto model-based semantically rich representations of network codes.

Assuming that the network codes could be digitalised to the same level as current digitalised representations, like Geographical Information System, and a federated Common Information Model (CIM) of the electrical grid was achieved it is conceivable that virtual mappings could be made between the nodes in the graph representation of the network codes and the CIM representations with the outputs of the mappings being used in load flow simulations to ingest the network code values relevant to the grid components that are partaking in the load flow analysis with any violations highlighted to the user. This would be a futuristic use case set in a fully digitalised electrical system utilising the regulatory and grid models in tandem to form part of a decision making system with minimal expert input.

All of the above future work items have the focus on building on a new representation of the network codes so that the divergence between the levels of digitalisation shared by the codes and the systems they govern is lessened in the hope that they can remain relevant and easily maintained in a rapidly changing environment.

The two previous future work items centre on the management of the network codes but there is also a case to provide consistency across the versions of the network codes in multi-actor scenarios. Section 4.2.3 proposed a new set of business logic steps and if consistency was to be realised these steps would need to be fully implemented. It is these mappings and a further step to smooth out the inconsistencies by term substitution to provide a federated set that could be shared and common codes managed by both affected actors. This proposed precision mapping system, once implemented and validated, could then be augmented by localisation to provide a pan regional capability and potentially a pan-European set, where all versions would share a common representation and maybe agree a consistent framework that would be built up of all the shared codes. This common representation would have the potential to allow codes to be updated in a federated way with the actor and regional varieties still maintained at the correct level. Indeed, the network codes would be stored in a graph representation, and each actor's local version would be extracted by means of a query. Also the output format of that query could take the form of a PDF document; eventually the shared graph representation would become definitive.

If all of these challenges were able to be resolved, the industry would be able to see the benefits of a shared semantically sound foundation for network, grid and distribution codes. At this point, industry players, regulators and other stakeholders might be open to more revolutionary concepts, such as

- replacing the current paper documents with linked ontologies and tools to author, update, and perform conflict detection and resolution;
- use of blockchain-backed smart contracts to underpin commercial relationships, from small suppliers to national providers;
- the code *documents* would be derived as needed from their computer representations, and would be informational only (non-normative);
- the same codes representation could be used internationally and could form the basis of interconnection agreements between countries;
- software would be able to search for discrepancies between supply and demand at all levels, and to perform static analysis of desirable properties like system-wide voltage stability;

- monitoring systems would be able to consult the computer representation of the codes and thereby to classify smart grid behaviour as anomalous or not.

Each of the above future work items builds a new representation of the network codes so that the divergence between the levels of digitalisation shared by the codes and the systems they govern is lessened, in the hope that they can remain relevant and unchallenged.

Appendix A

Tools and Technologies

To support the experiments described in Chapter 4, a set of tools and technologies were used and from them a system was built. Features were added to the system as needed by the research experiments, so development of this system was highly suited to using the agile methodology Beck et al. (2001).

The system offers a typical set of text mining operations: **Data Ingestion**, **Data Normalisation**, **Data Extraction**, **Concept Modelling** and **Query Interface**. Figure A.1 illustrates the high-level architecture of the system. In section A.4, the system architecture, its tools and techniques are described.

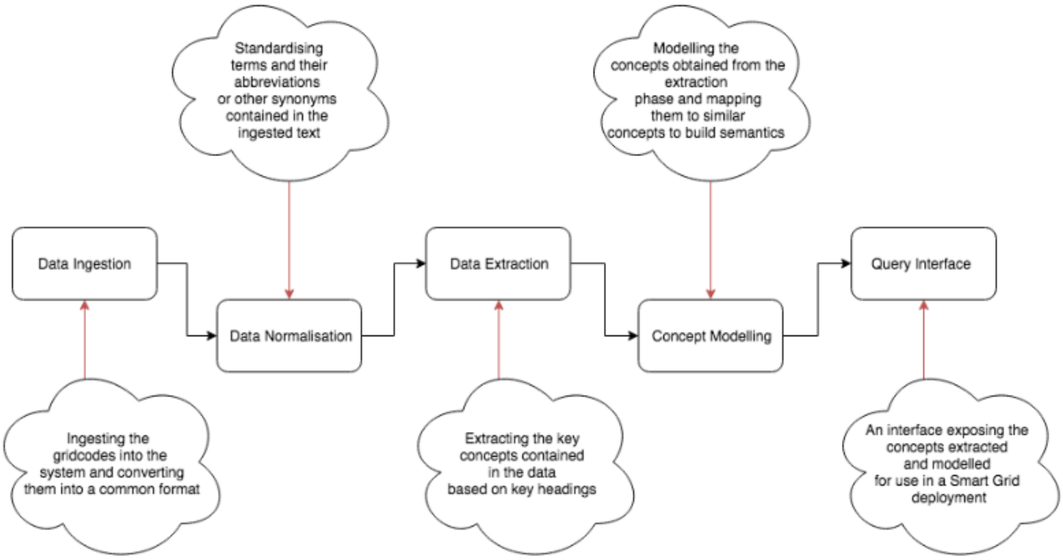

Standardising terms and their abbreviations or other synonyms contained in the ingested text
Modelling the concepts obtained from the extraction phase and mapping them to similar concepts to build semantics
Data Ingestion
Data Normalisation
Data Extraction
Concept Modelling
Query Interface
Ingesting the gridcodes into the system and converting them into a common format
Extracting the key concepts contained in the data based on key headings
An interface exposing the concepts extracted and modelled for use in a Smart Grid deployment

Fig. A.1 The Text Mining Process

The following sections detail the tools and technologies used to both host and perform the operations needed in the text mining process.

A.1 Text Ingestion Technologies

The aim of the text ingestion for the system was to perform three main tasks: a) the ingestion of the raw PDF documents into the system, b) the extraction of the text and metadata from the PDF source and c) the storage of the raw documents, the extracted text and the meta-data for text mining and access by other components. This component was built using the MEAN Stack Ihrig and Bretz (2015), which is a JavaScript based framework using MongoDB, Express, NodeJs and Angular to expose, create and present the relevant functionality needed to ingest, parse and store the data. The following subsections break down this subsystem into its component parts and discuss the techniques used and why they were chosen.

A.1.1 Ingestion

An interface is needed to take a PDF file from a desktop device and upload it to the system for processing. It was a design goal that the user should not be required to perform any preprocessing on the data and their sole goal in this ingestion process would be to simply identify the file for upload, so it was decided to use a browser-based method of upload using REST. Figure A.2, outlines the process of how a user interacts with the browser to send the file to the text ingestion service via a REST Post request so that the data is in the system and ready for processing.

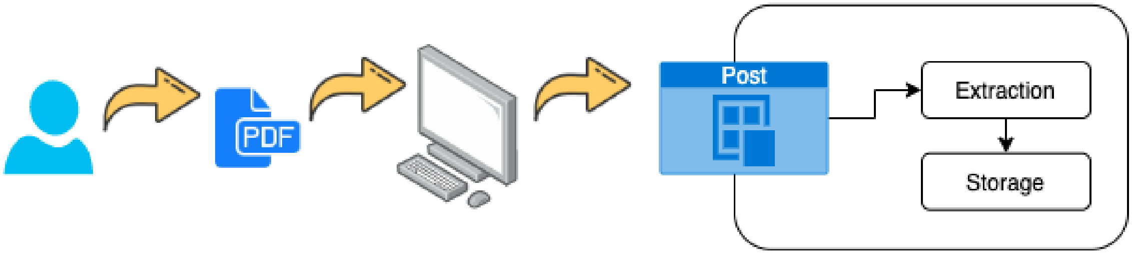


Fig. A.2 The Text Ingestion Process

A.1.2 Extraction

To extract the relevant data, as required for the experiments in chapter 4, three main data items are needed. They are, the extracted text, the metadata and the original document, so the process must extract the text and meta-data from the PDF while maintaining the structure and integrity of the original document. As can be seen in A.3, to perform this an open source NodeJS module, pdf-parse[1], was used and from that 2 distinct JavaScript Object Notation (JSON) objects were created. Other options were considered but pdf-parse performed particularly well.

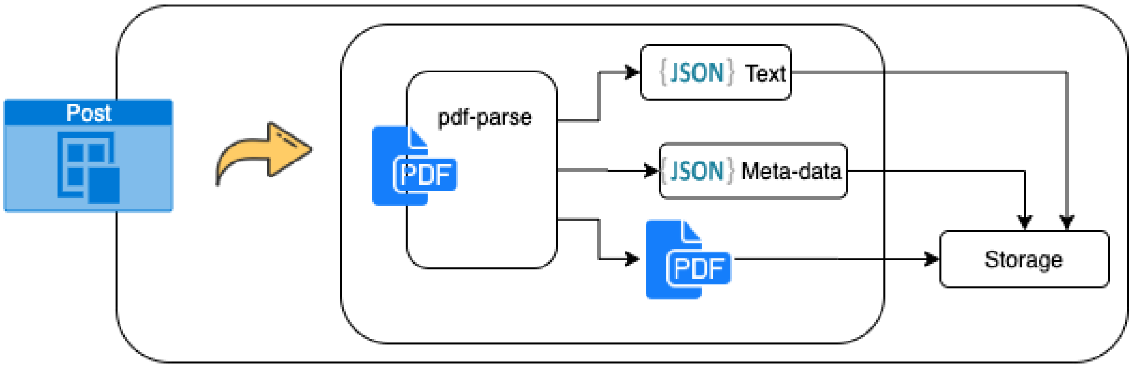


Fig. A.3 The Text Extraction Process

[1] https://gitlab.com/autokent/pdf-parse/-/blob/726214b1f055c8cb6d9421faf16ede0402e3a1d7/README.md

A.1.3 Storage

Persisting the extracted text, metadata and raw PDF file in a convenient fork is needed to support later operations. The extracted text and metadata are in JSON format and the raw files are in PDF, so the choice of storage technology needs to reflect this. Among other options, PostgreSQL[2] and MongoDB[3] were explored as potential object stores and both were found a viable solutions but given that MongoDB stores the data in BSON[4], a binary version of JSON, format it was decided that it was the preferable option. Furthermore, given that MongoDB is a component of the MEAN stack framework it has a native compatibility with NodeJs and its drivers. At this stage in the process the data is relatively unprocessed and contained in a large block of text so the use of a standard MongoDB collection would not be suitable. Therefore MongoDB's GridFS [5] storage option is used to store the data. This option stores the raw file, the data and metadata as a property of a file reference in the database, ensuring traceability is maintained.

[2] https://www.postgresql.org/

[3] https://www.mongodb.com

[4] http://bsonspec.org/

[5] https://docs.mongodb.com/manual/core/gridfs/

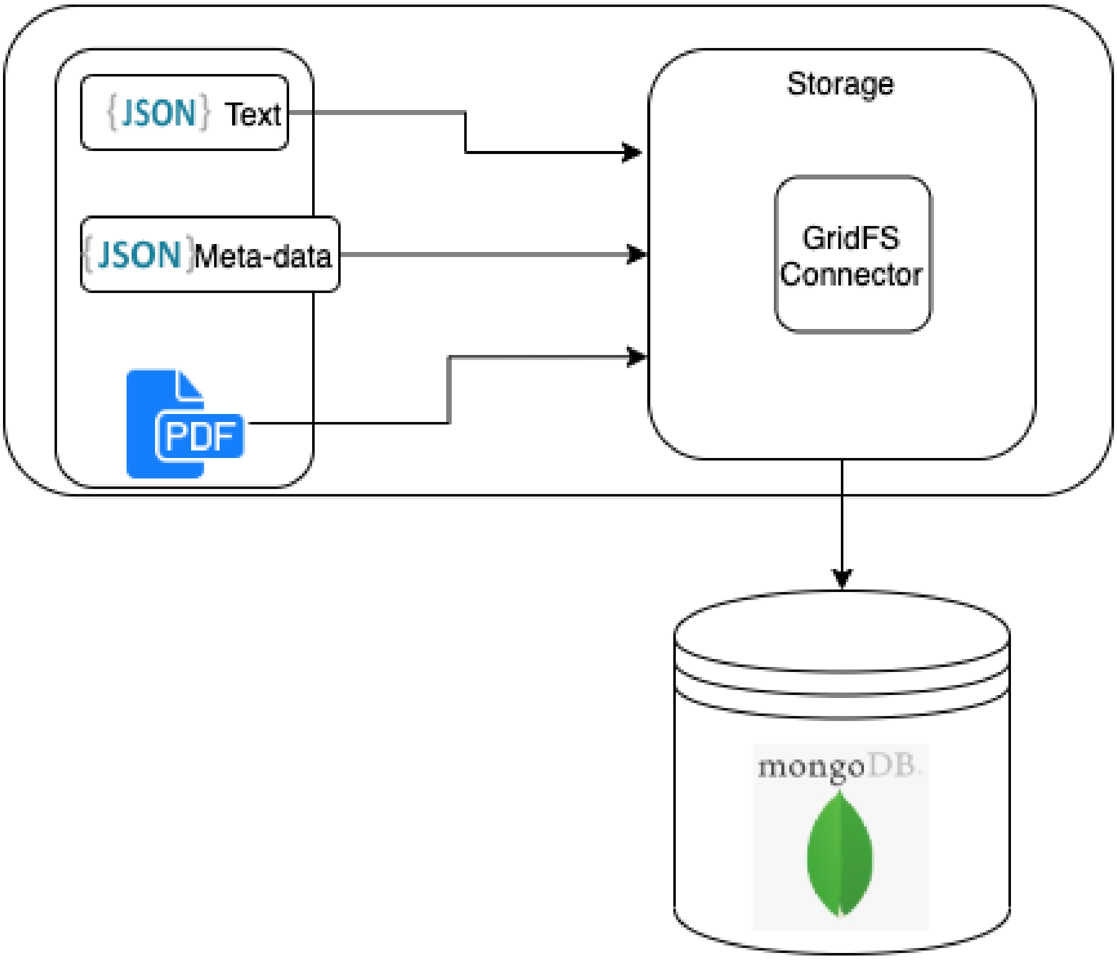


Fig. A.4 The Text Storage Process

A.2 Text mining Techniques and Technologies

At this stage in the process, the data has been ingested and the normalisation phase has began in that the data contained in the files are in a common computer readable format. The next phase completes the normalisation task, then cleans the data cleansing and extracts the concepts. This phase operates as a pipeline. For reasons of maturity and flexibility, the phase is based on tools from the Python ecosystem, notably Pandas [6]. The input into the pipeline will be the raw text with the output being concept objects that are ready to be modelled and stored. The following subsections detail this pipeline and its operations.

[6] https://pandas.pydata.org/

A.2.1 Data Cleansing

Given that the data, at this stage, is in a human readable text format containing punctuation, table of contents, footers and other unnecessary elements, prior to further processing it is necessary to remove such unwanted data. Figure A.5 illustrates this part of the pipeline and the cleansing comprises four clearly defined operations, the removal of the table of contents, the removal of all punctuation, the removal of footers and other unnecessary items and finally the normalisation of headings as labels for their text.

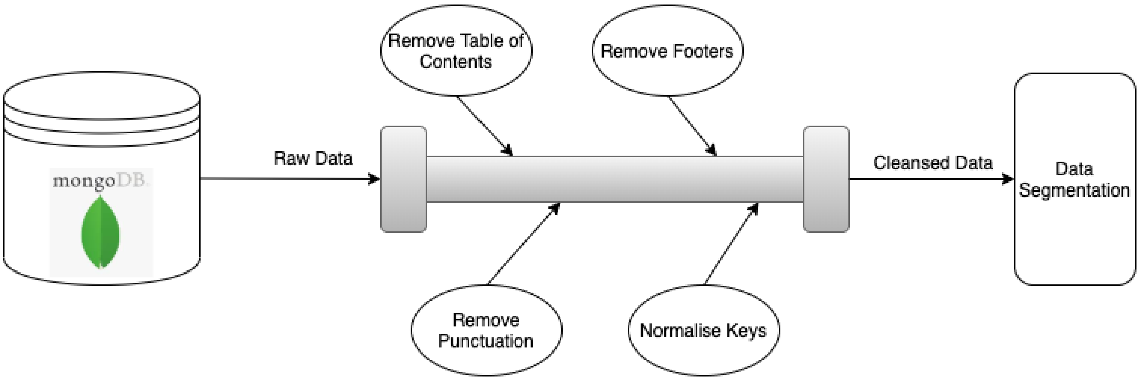


Fig. A.5 The Text Cleansing Process

In the beginning, the Table of Contents (TOC) was used for data segmentation to feed the Data Segmentation pipeline but this was found to be insufficient and so would not provide a good foundation for later steps. The decision was taken to remove the TOC to eliminate duplication and misleading data. Each line in the text to was passed to a function that analysed the line using regular expressions and the outcome would determine if it was a TOC item and, if so, it was removed. The next operation in the pipeline used `String.punctuation` to identify punctuation items like commas, full stops and question marks and to remove them from the text. The operation to remove footers uses regular expression matching, analogues to what was used to match and remove the TOC. The last operation in the data cleansing pipeline identifies key items, like headings for both the Annex and the main body, using regular expressions and a hard-coded dictionary of terms. The terms were normalised them by capitalising the text and removing trailing and leading spaces.

A.2.2 Data Segmentation

The grid code documents are structured hierarchically (with headings, subheadings and sub-subheadings etc.) so this structure needs to be captured and used because it provides essential context for any of the concepts in the document. The operations in the data segmentation pipeline are designed to recognise headings and subheadings, to segment the concepts so that their position in the (nested) document structure is captured and stored with each concept. Figure A.6 illustrates the three operations that identify the sections, derive the structure and add relationships to the sections.

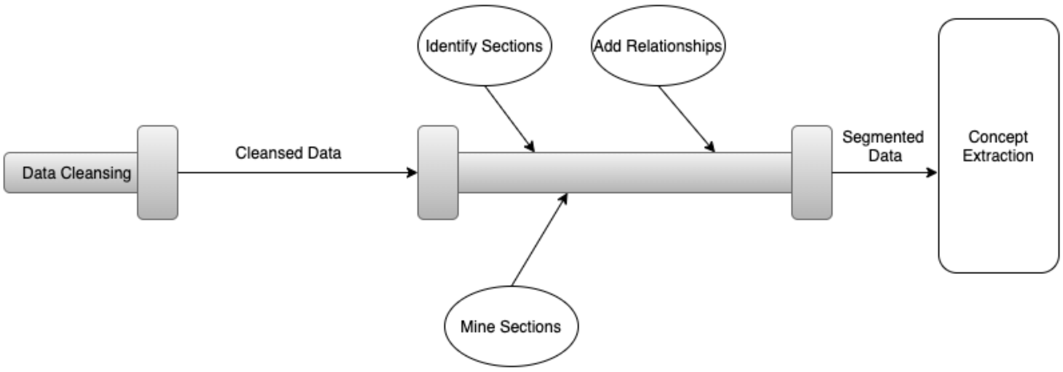


Fig. A.6 The Data Segmentation Process

The first operation identifies the section, its start line number and its end line number. In the network codes there are headings and keys and to help this we add a mapping the system can use to identify the keys. The keys at the beginning of each section or subsection are in an alphanumeric format similar to OCC112, which would stand for Operations Code 1.1.2, before the keys were normalised in the previous operation. The next operation uses the sections to create an object that contains the text, the key and the heading of the section or subsection. The next operation after this defines the relationships, by analysing the number after the key and determining if it is a section, subsection or further nested item and defines one of two relationships, **IS CATAGORISED AS** or **IS SUBSECTION OF**. These are used to build up a relationship map that mirrors the structure of the original codes but in a way that it can be represented in the concept mapping.

A.3 Concept Storage and Traversal

This phase of the operation takes the relationship map and the segmented text, storing this data so that it can be traversed easily. Three potential solutions were briefly explored: relational, noSQL and Graph databases, in terms of how each technology would facilitate storage, concept traversal and analysis. While all of these models could be used for storage, and either document-oriented or graph databases could be used for analysis, traversal operations are native only to graph databases and so this model was chosen. We chose Neo4J[7], due to the data and the relationship carrying equal weight in in the concept of the graph database. The data at this phase is segmented and the relationships are mapped in a way that matches the structure of the document in terms of how each of the sections and subsections are nested in a hierarchical way. Figure A.7 depicts the flow of the concept representation phase, using a series of inserts and updates into the graph database using the query language used in Neo4J called Cypher. This is a four step process with the first two converting the segmented data into a format that matches the format of a node and then extracting the relationship from the relationship map. The next two phases create each node and use the relevant relationship map items to create relationships between nodes.

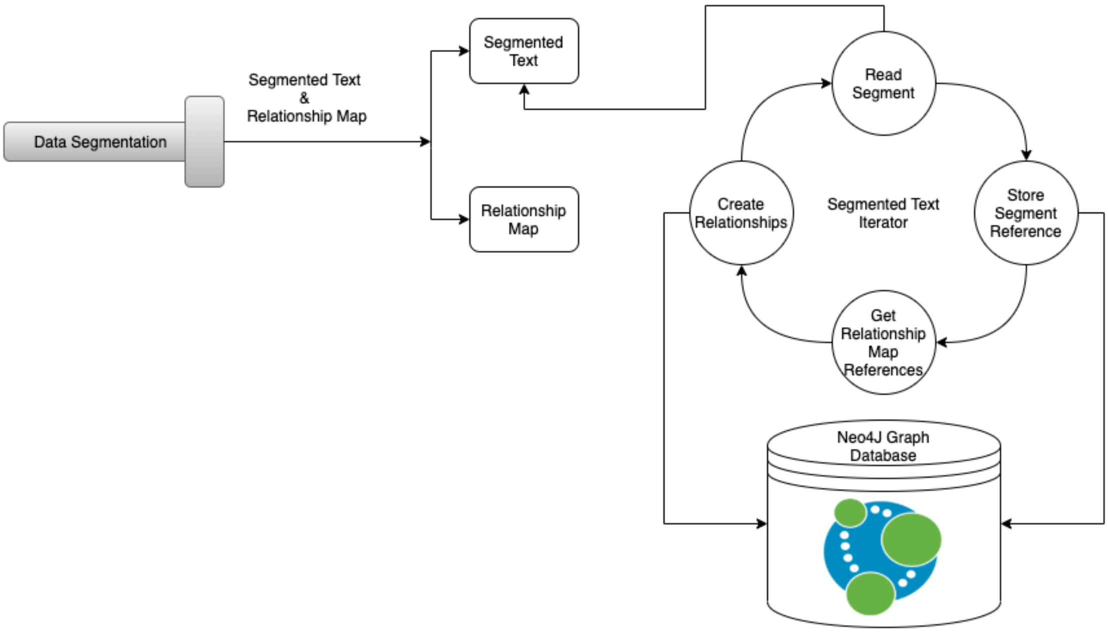


Fig. A.7 The Concept Representation Process

[7] https://neo4j.com/

A.4 System Architecture

The previous sections in this chapter define the pipeline and its technologies and processes. Referring to figure A.1, the process used is modular and the architecture should reflect this. The architecture also needs to be forward thinking and to have the potential to support future needs like validation and use cases that have not been specified yet. A microservices architecture [8] was chosen that allowed the separate stages of the process to be abstracted into modules that share resources and communicate with each other to share data and functionality. Kubernetes [9] was used to deploy the microservices in a scalable and easily-managed manner; it is a standard deployment method and is compatible with a host of widely used enterprise hosting platforms that are used by energy utilities.

[8] https://microservices.io/

[9] https://kubernetes.io/

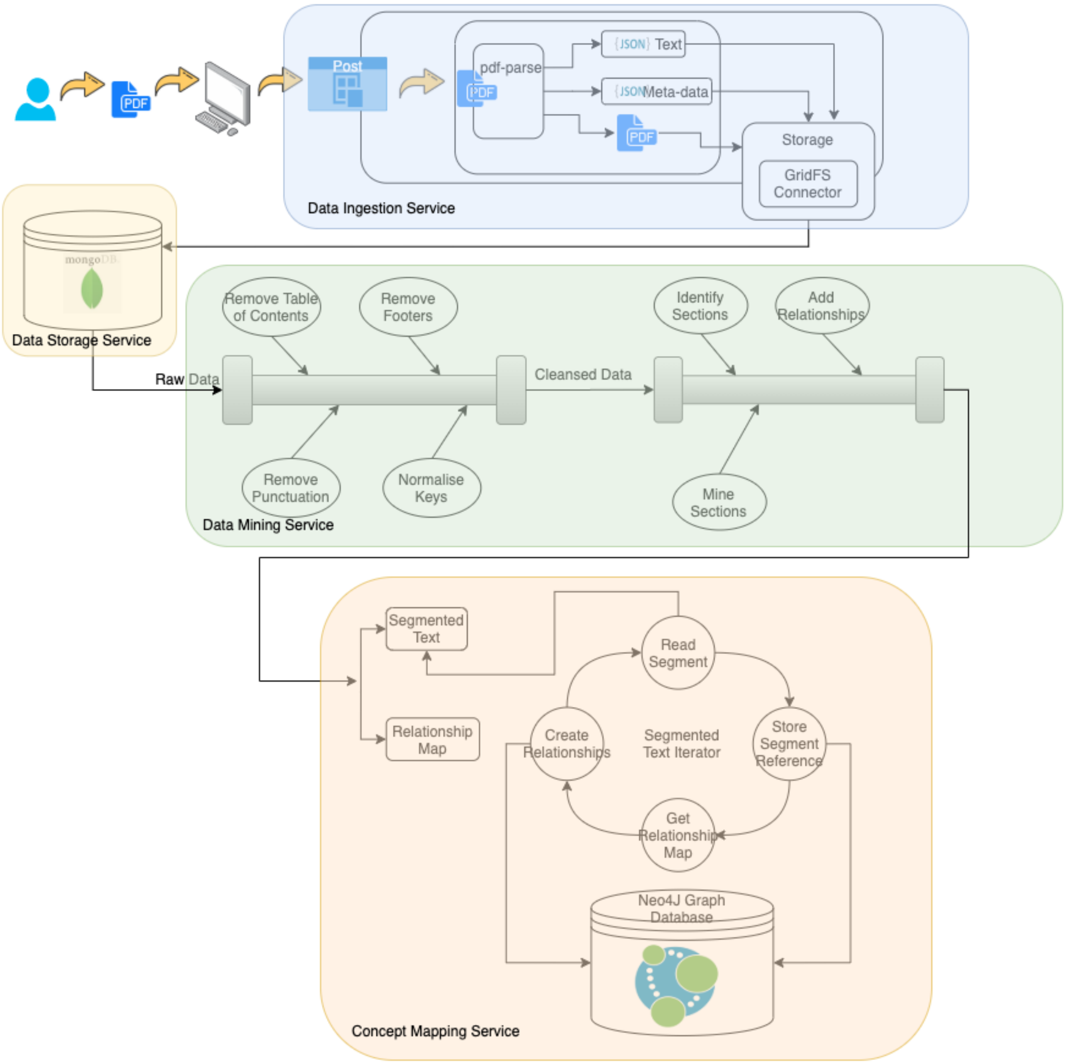


Fig. A.8 The Overall System Architecture

Bibliography

Allan, R. (2001). *A History of the Personal Computer: The People and the Technology*. Allan Pub.

Antonacopoulos, A. and Karatzas, D. (2004). Document image analysis for World War II personal records. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 336–341.

Anwar, A. and Mahmood, A. N. (2016). Anomaly detection in electric network database of smart grid: Graph matching approach. *Electric Power Systems Research*, 133:51 – 62.

Barbulescu, L., Watson, J.-P., and Whitley, L. D. (2000). Dynamic representations and escaping local optima: Improving genetic algorithms and local search. *AAAI/IAAI*, 2000:879–884.

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., and Thomas, D. (2001). Manifesto for agile software development.

Bienz, T., Cohn, R., and Adobe Systems (Mountain View, C. (1993). *Portable document format reference manual*. Citeseer.

Black, P. E. (2021). Jaro winkler. *Dictionary of Algorithms and Data Structures*.

Boyer, S. A. (2009). *Scada: Supervisory Control And Data Acquisition*. International Society of Automation, Research Triangle Park, NC, USA, 4th edition.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (2008). Extensible markup language (xml). https://www.w3.org/TR/REC-xml/.

Buckland, M. (1998). What is a digital document ? *Document numérique*, 2(2):221–230.

Bunke, H. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18(8):689 – 694.

Bunke, H. (1999). Error correcting graph matching: on the influence of the underlying cost function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):917–922.

Cason, T. P., Absil, P.-A., and Van Dooren, P. (2013). Iterative methods for low rank approximation of graph similarity matrices. *Linear Algebra and its Applications*, 438(4):1863–1882.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions.

Chakravarty, S., Dogan, K., and Tomlinson, N. (2006). A hedonic study of network effects in the market for word processing software. *Decision Support Systems*, 41(4):747 – 763. Economics and Information Systems.

Charlton, N., Greetham, D. V., and Singleton, C. (2013). Graph-based algorithms for comparison and prediction of household-level energy use profiles. In *2013 IEEE International Workshop on Inteligent Energy Systems (IWIES)*, pages 119–124.

Chen, W.-Y., Zhang, D., and Chang, E. Y. (2008). Combinational collaborative filtering for personalized community recommendation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 115–123, New York, NY, USA. Association for Computing Machinery.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and control*, 2(2):137–167.

Cinque, L., Yasuda, D., Shapiro, L., Tanimoto, S., and Allen, B. (1996). An improved algorithm for relational distance graph matching. *Pattern Recognition*, 29(2):349 – 359.

CONTE, D., FOGGIA, P., SANSONE, C., and VENTO, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298.

Crockford, D. (2006). The application/json media type for javascript object notation (json)," rfc 4627 (informational), internet engineering task force. http://www.ietf.org/rfc/rfc4627.txt.

Déjean, H. and Meunier, J.-L. (2006). A system for converting pdf documents into structured xml format. In *International Workshop on Document Analysis Systems*, pages 129–140. Springer.

Department of Communications, C. A. and the Environment (2020). Single electricity market (sem). https://www.dccae.gov.ie/en-ie/energy/topics/Electricity/commission-for-energy-regulation-(cer)/Pages/Single-Electricity-Market-(SEM).aspx.

Dori, D., Doermann, D., Shin, C., Haralick, R., Phillips, I., Buchman, M., and Ross, D. (1997). The representation of document structure: A generic object-process analysis. In *Handbook of character recognition and document image analysis*, pages 421–456. World Scientific.

E.DSO (2020). Why smart grids? https://www.edsoforsmartgrids.eu/home/why-smart-grids/.

Eirgrid (2020). Demand side unit (dsu). http://www.eirgridgroup.com/customer-and-industry/becoming-a-customer/demand-side-management/.

EirGrid Grid Code (2015). Grid Code v6 EirGrid Grid Code. http://www.eirgridgroup.com/site-files/library/EirGrid/GridCodeVersion6.pdf. Last accessed: 20181129.

EN50160 (1999). Voltage characteristics of electricity supplied by public distribution systems.

ENTSO-E (2020). Entso-e member companies. https://www.entsoe.eu/about/inside-entsoe/members/.

Erfianto, B., Mahmood, A. K., and Rahman, A. S. A. (2007). Modeling context and exchange format for context-aware computing. In *2007 5th Student Conference on Research and Development*, pages 1–5.

ESB (2016). Distribution code. https://www.esbnetworks.ie/docs/default-source/publications/distribution-code-v5-0.pdf. Last accessed: 20181129.

ESRI (2000). Escos in ireland: Investigation of energy service companies in 2000. https://www.esri.ie/publications/escos-in-ireland-investigation-of-energy-service-companies-in-2000.

Ferman, A. M., Errico, J. H., Beek, P. v., and Sezan, M. I. (2002). Content-based filtering and personalization using structured metadata. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 393–393.

for Energy Regulation, C. (2014). Rate of Change of Frequency (RoCoF) Modification to the Grid Code. https://www.cru.ie/wp-content/uploads/2014/07/CER14081-ROCOF-Decision-Paper-FINAL-FOR-PUBLICATION.pdf. Last accessed: 20181129.

Frankenfield, J. (2019). Business logic. https://www.investopedia.com/terms/b/businesslogic.asp.

Furuta, R. (1995). Documents in the Digital Library.

Gaffney, F., Deane, J., and Gallachóir, B. (2017). A 100 year review of electricity policy in ireland (1916–2015). *Energy Policy*, 105:67 – 79.

Gao, X., Zhu, Z., Hao, X., and Yu, H. (2017). An effective collaborative filtering algorithm based on adjusted user-item rating matrix. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 693–696.

Ghosh, D., Ghose, T., and Mohanta, D. (2013). Reliability analysis of a geographic information system-aided optimal phasor measurement unit location for smart grid operation. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 227(4):450–458.

Harary, F. and Palmer, E. M. (1973). Chapter 1 - labeled enumeration. In Harary, F. and Palmer, E. M., editors, *Graphical Enumeration*, pages 1 – 31. Academic Press.

Harpale, A. S. and Yang, Y. (2008). Personalized active learning for collaborative filtering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 91–98, New York, NY, USA. Association for Computing Machinery.

Ihrig, C. J. and Bretz, A. (2015). *Full Stack JavaScript Development With MEAN*. Sitepoint, 1st edition.

Jeh, G. and Widom, J. (2002). Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543.

Jiang, D. and Yang, X. (2009). Converting pdf to html approach based on text detection. In *Proceedings of the 2nd international conference on interaction sciences: Information technology, culture and human*, pages 982–985.

Johnson, D. S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256 – 278.

Jäger, A., Mittelstädt, S., Oelke, D., Sander, S., Platz, A., Bouwman, G., and Keim, D. A. (2016). Lessons on combining topology and geography : Visual analytics for electrical outage management. In Andrienko, L. and Sedlmair, M., editors, *EuroVis Workshop on Visual Analytics*. The Eurographics Association. Article Number: 1116.

Kanayama, F., Nishibayashi, Y., Yonezawa, Y., and Doi, Y. (2014). Development of autonomous power electronics products with communication middleware. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 61–66.

Kim, D., Alaerjan, A., Lu, L., Yang, H., and Jang, H. (2017). Toward interoperability of smart grids. *IEEE Communications Magazine*, 55(8):204–210.

Koch, E. (2014). Demand response management system. US Patent 8,782,190.

Koutra, D., Parikh, A., Ramdas, A., and Xiang, J. (2011). Algorithms for graph similarity and subgraph matching. In *Proc. Ecol. Inference Conf*, volume 17.

Kumar, Pradeep, Singh, and Asheesh (2012). *Renewable Energy Desalination*. The World Bank.

Li, D. and Dick, S. (2019). Residential household non-intrusive load monitoring via graph-based multi-label semi-supervised learning. *IEEE Transactions on Smart Grid*, 10(4):4615–4627.

Liang, J., Doermann, D., and Li, H. (2005). Camera-based analysis of text and documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2):84–104.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Lyons, P., Ryan, D., and Butler, B. (2018). Notes from a briefing meeting.

Melnik, S., Garcia-Molina, H., and Rahm, E. (2002). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In *Proceedings 18th International Conference on Data Engineering*, pages 117–128.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Elsevier Science & Technology, Saint Louis, UNITED STATES.

Neo4J (2021). Sorensen dice similarity. *APOC Documentation of Procedures and Functions for Text Operations*.

Networks, E. (2009). Conditions governing the connection and operation of micro-generation'. https://www.esbnetworks.ie/docs/default-source/publications/conditions-governing-the-connection-and-operation-of-micro-generation.pdf?sfvrsn=4.

Nikolić, M. (2012). Measuring similarity of graph nodes by neighbor matching. *Intelligent Data Analysis*, 16:865–878.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Nkgau, T. and Anderson, G. (2017). Graph similarity algorithm evaluation. In *2017 Computing Conference*, pages 272–278.

Norrie, M. and Signer, B. (2003). Switching over to paper: a new Web channel. In *Proceedings of the 7th International Conference on Properties and Applications of Dielectric Materials (Cat. No.03CH37417)*, pages 209–218. IEEE Comput. Soc.

Osburn III, D. C. (2003). Remote terminal unit. US Patent 6,628,992.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76–76.

Rashed Mohassel, R., Fung, A., Mohammadi, F., and Raahemifar, K. (2014). A survey on advanced metering infrastructure. *International Journal of Electrical Power and Energy Systems*, 63:473 – 484.

Rawat, D. B. and Bajracharya, C. (2015). Detection of false data injection attacks in smart grid communication systems. *IEEE Signal Processing Letters*, 22(10):1652–1656.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Ryan, D., Ponce De Leon, M., Grant, N., Butler, B., Vogel, S., Mirz, M., and Lyons, P. (2019). Deriving policies from connection codes to ensure ongoing voltage stability. *Energy Informatics*, 2(1):19.

Samitier, C. (2017). *IEC 61850 Communication Model*, pages 7–9. Springer International Publishing, Cham.

Santodomingo, R., Rohjans, S., Uslar, M., Rodríguez-Mondéjar, J., and Sanz-Bobi, M. (2014). Ontology matching system for future energy smart grids. *Engineering Applications of Artificial Intelligence*, 32:242 – 257.

SEAI (2013). Eligibility Criteria for Heating and Electricity Provision regarding Inverters. https://www.seai.ie/publications/Inverter.pdf. Last accessed: 20181129.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Song, W., Duan, Z., Yang, Z., Zhu, H., Zhang, M., and Tang, J. (2019). Explainable knowledge graph-based recommendation via deep reinforcement learning.

Specht, M. and Rohjans, S. (2013). *ICT and Energy Supply: IEC 61970/61968 Common Information Model*, pages 99–114. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tyson, H. (2007). *Microsoft Word 2007 Bible*. Bible. Wiley.

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42.

Wang, J., de Vries, A. P., and Reinders, M. J. T. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 501–508, New York, NY, USA. Association for Computing Machinery.

Wang, Z., Tan, Y., and Zhang, M. (2010). Graph-based recommendation on social networks. In *2010 12th International Asia-Pacific Web Conference*, pages 116–122.

Watts, D. (2004). *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton studies in complexity. Princeton University Press.

Zager, L. A. and Verghese, G. C. (2008a). Graph similarity scoring and matching. *Applied Mathematics Letters*, 21(1):86 – 94.

Zager, L. A. and Verghese, G. C. (2008b). Graph similarity scoring and matching. *Applied mathematics letters*, 21(1):86–94.

Zhou, F. and De la Torre, F. (2016). Factorized graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1774–1789.