A MODEL FOR THE VISUAL REPRESENTATION OF THE COHERENCE OF FACTS IN A TEXTUAL DOCUMENT SET

by

Louis Engelbrecht

Submitted in accordance with the requirements for
the degree of

MASTER OF SCIENCE

in the subject

COMPUTING

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor: Prof. Adele Botha

2016

DECLARATION

I declare that **A model for the visual representation of the coherence of facts in a textual document set** is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution.

_____ _____

SIGNATURE DATE

(Mr)

ACKNOWLEDGEMENTS

I would like to acknowledge the contribution of the following people to the successful completion of the research study:

- my research study leader, Prof. Adele Botha, who has spent many hours making sure that the study is on track. Without her help, assistance and advice, I would not have come this far;

- Marna Botha, for assisting with the statistical analysis of the feedback received from the solution demonstrations. Her help and assistance is greatly appreciated;

- the expert audience to whom the visualisation software was demonstrated. Thank you for your time, friendly reception and advice;

- my wife, Marisa, and our children Coert, Zani and Christo, for their support and understanding, for time spent on the study was time spent away from them;

- my colleagues at the CSIR, for their support and kind words, especially during the hard times;

- Lidia de Wet, for the language editing of my dissertation;

- the former inmates of the Mafeking concentration camp, who put their endurances in writing. Their testimonies were used as textual source set in the study;

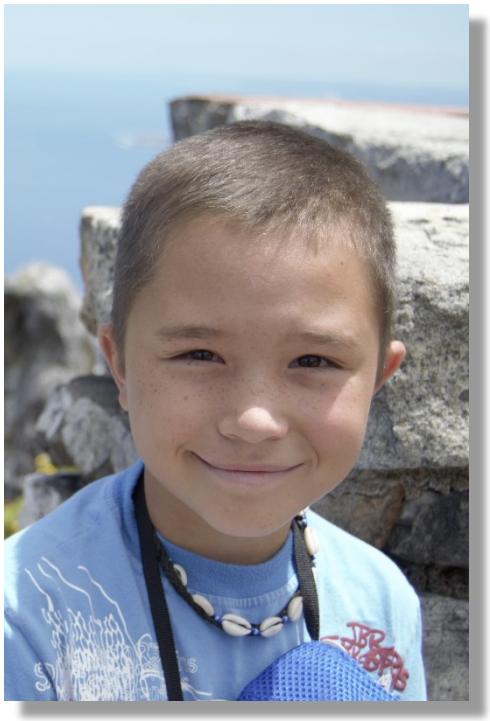
- lastly, I would like to thank God for giving me the opportunity and the strength to complete the research study:

Unless the Lord builds the house,
the builders labor in vain.

(Psalm 127)

DEDICATION

Dedicated to Marisa and our children, Coert, Zani and Christo.



In loving memory of my beloved son, Jan Abraham Christoffel (Christo) Engelbrecht

(3 May 2012 – 5 November 2015)

ABSTRACT

A large amount of information is contained in textual records, which originate from a variety of sources such as handwritten records and digital media like audio and video files. The information contained in these records is unstructured and to visualise the content of the records is not a trivial task. In order to visualise information contained in unstructured textual records, the information must be extracted from the records and transformed into a structured format.

This research aimed to visualise the coherence of facts contained in textual sources in order to allow the user who make use of the visualisation to make an assumption about the validity of the textual records as a set. For the purpose of the study, it was contemplated that the coherence of facts contained in a document set was indicated by the multiple occurrences of the same fact over several documents in the set.

The output of this research is a model that abstracts the process required to transform information contained in unstructured textual records into a structured format and the visual representation of the multiple occurrences of facts in order to support the process of making an assumption about the coherence of facts in the set. This assumption enables the user to make a decision, based on the coherence theory of truth, about the validity of the document set.

The model provides guidance and practices for performing tasks on similar textual document sets containing secondary data. The development of the model was informed by a phased construction of three specific software solution instantiations, namely an *initial information extraction*, an *intermediate visual representation* and a *final information visualisation* instantiation. The *final* solution instantiation was demonstrated to research participants and was evaluated as well. A pragmatic design science research approach was followed in order to solve the research problem. In conducting the research an adaption of the Peffers *et al*. (2006) design research process model was followed.

The result of the research is a model for the visual representation of the coherence of facts in a textual document set. Expert review of the model is added through a process of peer review and academic scrutiny by means of conference papers and a journal article. It is envisaged that the results of the research can be applied to a number of research fields such as Indigenous Knowledge, History and Law.

Key terms: Visualisation, Information Visualisation, Knowledge Visualisation, Data Visualisation, Document Visualisation, Design Science Research, Design Science Research Process Model, Information Extraction, Indigenous Knowledge, Coherence theory of truth.

LIST OF PUBLICATIONS

1. Engelbrecht, L., Botha, A., & Alberts, R. (2014). Information Visualisation View Design: Principles and Guidelines. In *International Conference on Computer Science, Computer Engineering, and Social Media* (pp. 17–28). Thessaloniki, Greece.

2. Engelbrecht, L. (2014). Master's research in progress: Visualisation of the coherence of facts contained in a textual record set. Presentation to FASTAR/Espresso Workshop 2014. Retrieved from www.fastar.org/publications/FEW 2014 - Engelbrecht.pdf

3. Engelbrecht, L., Botha, A., & Alberts, R. (2015). Designing the Visualization of Information. *International Journal of Image and Graphics*, *15*(2). http://doi.org/10.1142/S0219467815400057.

4. Engelbrecht, L., & Botha, A. (2015). Designing a View for Visually Representing Information Coherence in a Document set. In *9th IDIA conference, IDIA2015, Nungwi, Zanzibar* (pp. 177–188).

5. Engelbrecht, L., Botha, M., & Botha, A. (2016). Using Information Visualisation to give voice to a historical community. In *13th Prato CIRN Conference*. Prato, Italy.

ABBREVIATIONS

ABW	Anglo-Boer War
CPU	Central Processing Unit
CSS	Cascading Style Sheets
DOM	Document Object Model
DSR	Design Science Research
DST	Department of Science and Technology
HTML	Hypertext Markup Language
IE	Information Extraction
IP	Intellectual Property
ISO	International Organisation for Standardisation
IT	Information Technology
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
NARS	National Archives and Records Service
NER	Named-Entity Recognition
NIKMAS	National Indigenous Knowledge Management System
NRS	National Recordal System
PHP	Hypertext Preprocessor
SVG	Scalable Vector Graphics
SQL	Standard Query Language
TASC	Textual Analysis of Similar Crimes
XML	Extensible Markup Language

TABLE OF CONTENTS

DECLARATION ..... **i**

ACKNOWLEDGEMENTS **1**

DEDICATION **2**

ABSTRACT **3**

LIST OF PUBLICATIONS **4**

ABBREVIATIONS **5**

TABLE OF CONTENTS **6**

LIST OF FIGURES **10**

LIST OF TABLES **12**

CHAPTER 1: INTRODUCTION **14**

1.1 BACKGROUND *14*

1.2 RESEARCH PURPOSE AND RESEARCH QUESTIONS *14*

1.2.1 Rationale for research 15

1.2.2 Main research question 18

1.2.3 Subordinate research questions 18

1.2.4 Research purpose 19

1.2.5 Research objectives 19

1.3 RESEARCH DESIGN AND METHODOLOGY *21*

1.4 IMPORTANCE OF THE RESEARCH STUDY *24*

1.5 DELINEATION AND ASSUMPTIONS *24*

1.6 ETHICAL CONSIDERATIONS *25*

1.7 CHAPTER OUTLINE *26*

CHAPTER 2: RESEARCH APPROACH AND METHODOLOGY **28**

2.1 RESEARCH PARADIGM *28*

2.2 RESEARCH METHODOLOGY *30*

2.2.1 Design science research 32

2.2.1.1 Philosophical views of design science research 33
2.2.1.2 Artifacts produced by design science research 34
2.2.1.3 Knowledge contributions made by design science research 36
2.2.1.4 Guidelines for conducting design science research 38
2.2.1.5 Design science research cycles 39
2.2.1.6 Advantages and disadvantages of design science research 43

2.3 RESEARCH METHOD 44

2.4 RELIABILITY AND VALIDITY 51

CHAPTER 3: DEVELOPMENT OF THE INITIAL MODEL – TEXTUAL INFORMATION EXTRACTION 53

3.1 LITERATURE RELEVANT TO THE EXTRACTION OF INFORMATION FROM TEXTUAL SOURCES 54
3.1.1 Preprocessing tasks on textual sources prior to information extraction 55
3.1.2 Information extraction methods and tools 57
3.1.2.1 Information extraction methods 57
3.1.2.2 Information extraction tools 62
3.1.3 Storage of the extracted information is considered 64
3.1.4 Summary 67

3.2 A MODEL FOR THE EXTRACTION OF INFORMATION FROM A TEXTUAL DOCUMENT SET 68
3.2.1 PREPARE TEXTUAL DOCUMENTS FOR INFORMATION EXTRACTION 68
3.2.2 EXTRACT THE INFORMATION CONTAINED IN A TEXTUAL DOCUMENT SET 70
3.2.3 POPULATE THE DATABASE SCHEMA 73
3.2.4 IDENTIFY POSSIBLE ERRORS INTRODUCED DURING THE INFORMATION EXTRACTION PROCESS 75

3.3 SUMMARY 76

CHAPTER 4: DEVELOPMENT OF THE INTERMEDIATE MODEL – INFORMATION VISUALISATION 79

4.1 LITERATURE RELEVANT TO THE VISUALISATION OF INFORMATION COHERENCE 80
4.1.1 Similar work 80
4.1.2 Defining Visualisation, Knowledge Visualisation, Information Visualisation and Document Visualisation 82
4.1.2.1 Working definition of *Visualisation* 83
4.1.2.2 Working definitions of *Knowledge* and *Information* 85
4.1.2.3 Working definitions of *Information Visualisation* and *Knowledge Visualisation* 89
4.1.2.4 Document visualisation 91
4.1.2.5 Summary 93
4.1.3 Information Visualisation design principles and guidelines 93
4.1.3.1 Information Visualisation design principles and guidelines related to user interactions and interfaces 94

4.1.3.2 Information Visualisation design principles and guidelines related to additional information required by the user 96
4.1.3.3 Information Visualisation design principles and guidelines related to the profile of the user 97
4.1.3.4 Information Visualisation design principles and guidelines corresponding to the Eppler and Burkhard (2007) knowledge visualisation framework 98
4.1.3.5 Information Visualisation design principles and guidelines related to the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation 100
4.1.3.6 Information Visualisation design principles and guidelines related to the Elmqvist and Fekete's (2010) guidelines for implementing visual hierarchical aggregates 102
4.1.3.7 Information Visualisation design principles and guidelines related to the shortcomings of visualisations 103
4.1.4 Information Visualisation tools and techniques 106
4.1.4.1 Literature review of Visualisation techniques 106
4.1.4.2 Literature review of Visualisation tools 110
4.1.5 Summary 113
4.2 A MODEL FOR THE VISUAL REPRESENTATION OF THE COHERENCE OF INFORMATION IN A RECORD SET 113
4.2.1 Proposed design to visually represent the coherence of information in a record set 114
4.2.1.1 Design of the user interface 114
4.2.1.1.1 The overall view 118
4.2.1.1.2 The zoom-into-document view 120
4.2.1.1.3 The zoom-into-fact view 122
4.2.1.1.4 The domain information management and document management pages 124
4.2.1.1.5 Applying information visualisation principles and guidelines to the design of the visualisation view 128
4.2.1.2 Design of the database schema 136
4.2.2 Building a visualisation interface from the design 138
4.2.2.1 Technologies used in the implementation of the design 139
4.2.2.2 Screenshots (with mockup data) of the developed interface 141
4.3 SUMMARY 148
CHAPTER 5: DEVELOPMENT, DEMONSTRATION AND EVALUATION OF THE FINAL SOFTWARE SOLUTION AND VALIDATION OF THE FINAL MODEL 150
5.1 FINAL SOFTWARE SOLUTION 151
5.1.1 Final software solution demonstration 151
5.1.2 Final software solution evaluation 155
5.1.2.1 Feedback statements with regard whether the solution contributes towards identifying process components 156

5.1.2.2 Feedback statements with regard to the application of design principles and guidelines 158
5.1.2.3 Feedback statement with regard to usefulness of the solution artifact 169
5.1.2.4 Feedback comments and recommendations 171

5.2 THE FINAL MODEL 173
5.2.1 FINAL MODEL DEMONSTRATION 178
5.2.2 FINAL MODEL VALIDATION 180
5.2.2.1 Was the purpose of the study achieved in the model? 181
5.2.2.2 Does the model contains components to extract information from unstructured records? 182
5.2.2.3 Does the model contains visual representation components? 183
5.2.2.4 Was the development of the model informed by the construction of the instantiated software solution? 185

CHAPTER 6: CONCLUSION 187

6.1 SUMMARY OF CHAPTERS 187

6.2 OVERVIEW OF THE STUDY 188
6.2.1 The research process 188
6.2.2 Reflection on key findings 189

6.3 SIGNIFICANCE AND CONTRIBUTION OF THE RESEARCH 195

6.4 DELINEATION AND ASSUMPTIONS OF THE RESEARCH STUDY 197

6.5 REFLECTIONS 198
6.5.1 Scientific reflection 198
6.5.2 Methodological reflection 198
6.5.3 Personal reflection 200

6.6 FUTURE RESEARCH 201

6.7 CONCLUDING REMARKS 202

REFERENCES 203

APPENDIX A: PARTICIPANT CONSENT FORM 216

APPENDIX B: ETHICAL CLEARANCE CERTIFICATE 219

APPENDIX C: BACKGROUND INFORMATION ON THE HUYSER COLLECTION AND THE ANGLO-BOER WAR 220

LIST OF FIGURES

Figure 1.1: Models developed during the research phases 17
Figure 1.2: Research objectives and phases 20
Figure 1.3: Adapted design science research process from Peffers *et al*. (2006) 22
Figure 1.4: Chapter outline 27
Figure 2.1: DSR knowledge contribution framework (Gregor & Hevner, 2013) 37
Figure 2.2: Design science research cycles (Drechsler and Hevner, 2016) 40
Figure 2.3: Design science research process (Peffers *et al*., 2006) 45
Figure 3.1: Current phase in research – textual information extraction 53
Figure 3.2: Preliminary information extraction process 55
Figure 3.3: Proposed main components of the information extraction process 57
Figure 3.4: Word cloud example from Grant *et al*. (2015) 60
Figure 3.5: Example of a photographed page 69
Figure 3.6: Selecting factual information using ATLAS.ti 72
Figure 3.7: Snippet of XML data representation created by ATLAS.ti 73
Figure 3.8: Code snippet of *loadDocs* utility 74
Figure 3.9: Code snippet of texttodb.php script file 74
Figure 3.10: Code snippet of loadInformation utility 75
Figure 3.11: Initial model – extracting facts and transforming unstructured textual data into a structured form 77
Figure 4.1: Current phase in research – visual representation 79
Figure 4.2: Hierarchy of human understanding (Ackoff, 1989) 86
Figure 4.3: Preliminary Information Visualisation process 91
Figure 4.4: Proposed main components of an Information Visualisation solution 97
Figure 4.5: The Eppler and Burkhard (2007) knowledge visualisation framework 99
Figure 4.6: Circular treemap example provided by Viégas *et al*. (2013) 109
Figure 4.7: Design of the interaction between the visualisation views 115
Figure 4.8: Design of the overall view 118
Figure 4.9: Design of the zoom-into-document view 121
Figure 4.10: Design of the zoom-into-fact view 122
Figure 4.11: Displaying an occurrence of a fact in context 123
Figure 4.12: Displaying the contents of a document 124
Figure 4.13: Add/Edit domain information page design 125
Figure 4.14: Fact list page design 125
Figure 4.15: Document information Add/Edit page design 126
Figure 4.16: Document list page design 127

Figure 4.17: Zoomable circle-packing technique (Bostock, 2013b) 132
Figure 4.18: Database schema design 137
Figure 4.19: Three-tier application architecture 139
Figure 4.20: Technologies used in each layer 140
Figure 4.21: Starting page 142
Figure 4.22: Main visualisation interface 143
Figure 4.23: Zoom-into-fact view 143
Figure 4.24: Show in context dialog 144
Figure 4.25: Zoom-into-document view 144
Figure 4.26: Domain information management page 145
Figure 4.27: Add/Edit domain information dialog view 145
Figure 4.28: Document information management page 146
Figure 4.29: Add/Edit document dialog view 146
Figure 4.30: Coherence of the "measles occurred in camp" fact in the document set 147
Figure 4.31: Intermediate model - visually representing the coherence of facts in a structured data form 148
Figure 5.1: Current phase in research – final Information Visualisation model 150
Figure 5.2: Visualisation contains extracted information 157
Figure 5.3: Coherence of facts is represented 158
Figure 5.4: Visualising technique appropriate 159
Figure 5.5: Visualisation technique employs a minimal set of actions to complete a task 160
Figure 5.6: Functionality to control the level of detail 161
Figure 5.7: Visual layout is effective 162
Figure 5.8: Cognitive load is minimal 163
Figure 5.9: There is too much information on the visualisation 164
Figure 5.10: Additional functionality should be considered 165
Figure 5.11: Visualisation is clean and simple 166
Figure 5.12: Testimonies and facts are easily distinguishable from each other 167
Figure 5.13: Visualisation leads to misleading interpretations 168
Figure 5.14: Visualisation is interpretable 169
Figure 5.15: Useful for obtaining insights 170
Figure 5.16: Final model - resulting Information Visualisation model 174
Figure 5.17: Purpose of the study achieved in the model 182
Figure 5.18: Model contains information extraction components 183
Figure 5.19: Model contains visual representation components 184
Figure 5.20: Model informed by instantiton of the software solution 185
Figure 6.1: A model for visually representing the coherence of facts in a textual document set 194

LIST OF TABLES

Table 2.1: *Epistemology* and *ontology* of *Positivism*, *Interpretivism*, *Critical research* and *Pragmatism* as provided by Wahyumi (2010) 30
Table 2.2: Design science research artifacts 35
Table 2.3: Design science research guidelines (Hevner *et al*., 2004) 38
Table 2.4: Applying Hevner et al.'s (2004) design evaluation methods 49
Table 3.1: Example list of information extraction functions and tools 62
Table 3.2: Comparison between database storage systems (Agarwal & Srivastava, 2016) 66
Table 3.3: Mapping of study concepts to ATLAS.ti concepts 71
Table 3.4: Number of documents, unique facts and the number of times facts occurred 73
Table 4.1: Deriving Information Visualisation and Knowledge Visualisation 90
Table 4.2: Summary of working definitions of visualisation concepts 93
Table 4.3: Proposed Information Visualisation design principles and guidelines 104
Table 4.4: General Visualisation techniques 106
Table 4.5: Various visualisation software tools 110
Table 4.6: Roles of layers in the three-tier architecture 139
Table 5.1: Final solution demonstration participant profiles 151
Table 5.2: Final solution demonstration feedback questionnaire 153
Table 5.3: Contributions of feedback statements to the evaluation of the final solution 154
Table 5.4: The visualisation contains extracted information 156
Table 5.5: The coherence of facts contained in the Huyser collection is represented in the visualisation 157
Table 5.6: The technique for visualising the facts contained in the testimonies is appropriate 159
Table 5.7: The visualisation technique employs a minimal set of actions to accomplish a task 160
Table 5.8: Functionality to control the level of details is available 161
Table 5.9: The visual layout of the visualisaiton is effective 162
Table 5.10: The cognitive load on the user is minimal 163
Table 5.11: There is too much information on the visualisation view 164
Table 5.12: Additional functionality to show less testimonies and facts should be considered 165
Table 5.13: The visualisation is clean and simple 166
Table 5.14: Testimonies/documents and facts are easily distinguishable from each other 167
Table 5.15: The visualisation leads to misleading interpretations 168
Table 5.16: The visualisation is interpretable 169
Table 5.17: The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history 170
Table 5.18: Model demonstration feedback questionnaire 179
Table 5.19: Contributions of feedback statements to model validation 180
Table 5.20: Purpose of the study achieved in the model 181

Table 5.21: Model contains information extraction components 183
Table 5.22: Model contains visual representation components.....184
Table 5.23: Model informed by instantiton of the software solution.....185
Table 6.1: Summary of the research process 190
Table 6.2: Model components for the visual representation of the coherence of facts in a textual document set...195

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

The majority of the world's information is contained in unstructured textual documents that describe information in a natural language like English (McCallum, 2005). Some of the domains described by unstructured textual documents include the Medical health domain (e.g. medical patient records), History domain (e.g. testimonies on a historical event) and the Indigenous Knowledge domain (e.g. instructions on preparing traditional medicine). The information contained in these documents is considered to be secondary data. *Secondary data*, according to Kamins (1993, p. 1), is data that is “collected by others and archived in some form”.

According to McCallum (2005), the processing of unstructured textual data is limited to searching for text in the data. In order to enable detailed processing of unstructured textual data, it needs to be transformed into a structured form.

Chen, Ebert, Hagen, Robert, Shneiderman and Tory (2009) and Dove and Jones (2012) state that the purpose of data visualisation is to gain insight into data. Dove and Jones (2012, p. 2) regard the insight as “something that is gained” and as “something that is experienced”. Since both definitions of *data* and *information* are ambiguous (Keller & Tergan, 2005; Liew, 2013; Nurnberger, Seising, & Wenzel, 2009), the defined purpose of data visualisation can be applied to the visualisation of information as well.

The purpose of this research is to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set. It is argued by this research that visually representing the coherence of facts contained in a document set would be a first step towards making an assumption about the validity of the facts contained in the set. For the purpose of the study, it is proposed that the coherence of the facts in a document set is indicated by the multiple occurrences of the facts contained in multiple documents, and that visually representing these coherences would enhance the insight into the validity of the facts.

After having presented background information and the aim of the research study, the next section discusses the problem statement and research questions of the study.

1.2 RESEARCH PURPOSE AND RESEARCH QUESTIONS

The rationale for research, research question, research purpose and the significance of the research study are provided in this section of the dissertation.

1.2.1 Rationale for research

The National Recordal System (NRS) is an initiative by the South African Department of Science and Technology (DST) to capture and preserve Indigenous Knowledge. As part of this initiative, the National Indigenous Knowledge Management System (NIKMAS) was designed and developed (Fogwill, Viviers, Engelbrecht, Krause, & Alberts, 2011).

During the development of the system, it became apparent to the researcher that the visualisation of the textual content contained in Indigenous Knowledge, for instance, the preparation method of traditional medicine, was complex and difficult to achieve. In addition, it was found that certain elements were consistently found in several descriptions of Indigenous Knowledge and that the visual representation of these consistencies would assist a person in making an assumption about the validity of the knowledge being described.

This research study proposes that having a model for extracting factual information contained in textual documents into a structured format and the visual representation of the coherence of the factual information in the records would be a step towards making an assumption about the possible truth value of the document set as a whole. The assumption that the coherence of facts in a document set indicates some value of truth is based on the coherence theory of truth (Young, 2013).

According to Young (2013, p. 1), the coherence theory of truth states that “the truth of any (true) proposition consists in its coherence with some specified set of propositions”. For the purpose of the research study, the coherence theory of truth is applied as follows: The possible truth of any fact occurrence consists in its coherence with occurrences of the fact in a document set.

Roche (2013), in addition, proposes that the coherence of a proposition with other propositions is an indication of an increase in the probability that some truth is contained by the proposition. The proposition of Roche (2013) is, for the purpose of the study, applied as follows: *The coherence of a fact occurrence with other occurrences of the fact is an indication of an “increase” in the probability that some truth is contained by the fact.*

Based on the coherence theory of truth as described by Young (2013) and the proposition of Roche (2013), the coherence between occurrences of a fact in a document set is considered to be an indication of the possibility that the fact may contain some truth or not.

In the process of developing a model to visually represent the coherence of facts in unstructured textual documents, a textual data source had to be used as input. Although the purpose of this research study originated from the development of an Indigenous Knowledge capturing and protection system, Indigenous Knowledge records could not be used as data source. Intellectual property (IP) issues related to Indigenous Knowledge (Fogwill *et al*., 2011) led to the decision to use an alternative data source.

As a result, an alternative source of data, which exhibited the same characteristics as the original intended Indigenous Knowledge records, was identified and used. A data set has to exhibit the following characteristics that are inherent to Indigenous Knowledge to be considered as an alternative:

- The data source should consist of data that is either in the form of documents (e.g. written testimonies) or data that can be transformed into documents (e.g. oral story tellings of Indigenous Knowledge), where each part of the set describes facts about the same domain.

- The data should contain only a finite set (a number of not more than 50 documents is proposed as appropriate) of original or transformed documents.

- The data in the original or transformed document set should be textual and unstructured.

- The information contained in the documents should be secondary, i.e. information collected by someone who is not the originator of the information and the information should be archived in some form.

- The validity of the information described in the records should be open for scrutiny.

Instead of using Indigenous Knowledge records, the Huyser collection[1] of Anglo-Boer War testimonies (Huyser, n.d.) is used for the study because the collection does not have IP restrictions, is easily available, has already been transcribed and digitised by the researcher before the study commenced and in particular satisfies the above mentioned criteria to be considered as an alternative data source:

- Each document in the Huyser collection describes the experiences of Boer women and children during the Anglo-Boer War. The testimonies contained in each record were collected by H.W. Huyser. Each testimony is hand written and is currently housed at the National Archives and Records Service (NARS) of South Africa.

- The Huyser collection contain is comprised of a finite set of 37 documents.

- Each testimony in the set of Huyser documents is in a textual format and contains unstructured data.

- The information is collected by the researcher and is archived electronically.

- The validity of the testimonies is under scrutiny (Mafeking, n.d., para 22).

The research study provides a model for supporting the process of making an assumption about the validity of facts contained in a set of textual documents by visually representing

[1] Background information on the Huyser collection and the Anglo-Boer War is provided in Appendix C.

the coherence of the facts. This developed model is applicable to any data set which exhibits the characteristics required by a data source to be considered as an alternative source. Expert review of the model is added through a process of peer review and academic scrutiny by means of conference papers (Engelbrecht, Botha, & Alberts, 2014; Engelbrecht, Botha, & Botha, 2016; Engelbrecht & Botha, 2015) and a journal article (Engelbrecht, Botha, & Alberts, 2015).

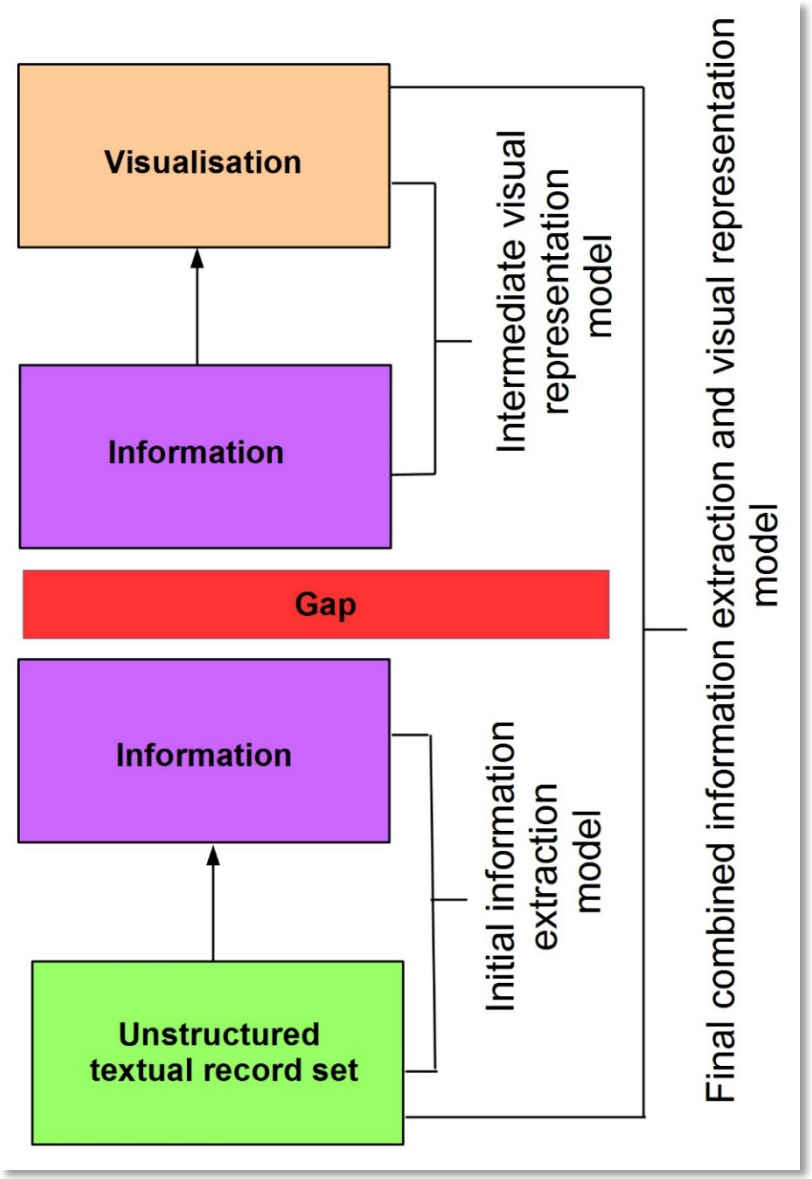


Figure 1.1: Models developed during the research phases

The proposed model is developed in three phases (depicted in Figure 1.1), namely the *initial* extraction of the information phase, the *intermediate* visual representation of the fact coherence phase and the *final* resulting phase. During each phase a specific software solution is designed and instantiated, which informed the construction of a model for that phase. During the *final* phase a model is defined for the visual representation of the

coherence of facts in a textual document set by combining the models from the *initial* information extraction and *intermediate* visual representation phases. It is therefore stated that the general model is informed by the specific software solution instantiation.

According to Keim, Rossi, Seidl, Verleysen and Wrobel (2012) and Mooranian, Dillon and Chang (2011), there exists a gap between the research fields of information extraction and visualisation. To progress towards visually representing the coherence of facts in a textual document set and to contribute to understanding the gap between the study fields of information extraction and visualisation, models as depicted in Figure 1.1 were developed during three phases:

- an information extraction model during the *initial* phase (see Chapter 3);
- an visual representation model during the *intermediate* phase (see Chapter 4);
- a model for the visual representation of the coherence of facts in a textual document set during the *final* phase. The *fina*l model is a conglomeration of the models defined by the previous phases (see Chapter 5).

In the next section, the research question that guides the study is discussed.

1.2.2 Main research question

The main research question of the study is:

> *What components can be combined into a model to transform unstructured textual data and to visually represent the coherence of facts contained in the set?*

More research questions are required to answer the main research question. The next section provides the required subordinate research questions.

1.2.3 Subordinate research questions

In order to answer the main research question, the following subordinate research questions have to be answered:

1. What components can be combined into a model to extract facts and transform unstructured textual data into a structured form?
2. What visualisation design principles can be applied to guide the design of an information visualisation interface?
3. What components can be combined into a model to visually represent the coherence of facts contained in a structured data form?

Subordinate research questions 1 and 3 are answered through a design science research (DSR) approach, which is discussed in Chapter 2. In order to answer the research question

2, Section 4.1.3 identifies visualisation design principles identified by means of a literature review.

The objectives of the research study are derived from the research questions. In the next section, the main research objective of the research study will be discussed.

1.2.4 Research purpose

The research purpose of the study corresponds to the main research question as defined in Section 1.2.2.

The purpose of the research is

> *To develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set.*

1.2.5 Research objectives

In conducting the research, the study considers the following research objectives:

1. to identify studies similar to this research;
2. to identify tools and techniques that can be used to extract facts from structured textual data;
3. to identify components that can be combined to extract facts and transform unstructured textual data into a structured form;
4. to identify visualisation design principles that can be applied to the design of an visual representation interface;
5. to identify tools and techniques to visually represent the coherence of facts in a structured form;
6. to identify components that can be combined to visually represent the coherence of facts contained in a structured data form;
7. to design, develop and instantiate a specific *final* software solution by combining the components of specific *initial* information extraction and *intermediate* visual representation software solution instantiations;
8. to evaluate the final software solution;
9. to define a general model for the visual representation of the coherence of facts in a textual document set by combining the components of the specific information extraction and visual representation software solutions;

10. to validate the research results, i.e. the model for the visual representation of the coherence of facts in a textual document set.

The objectives of the study and the research phases addressing them are depicted in Figure 1.2.

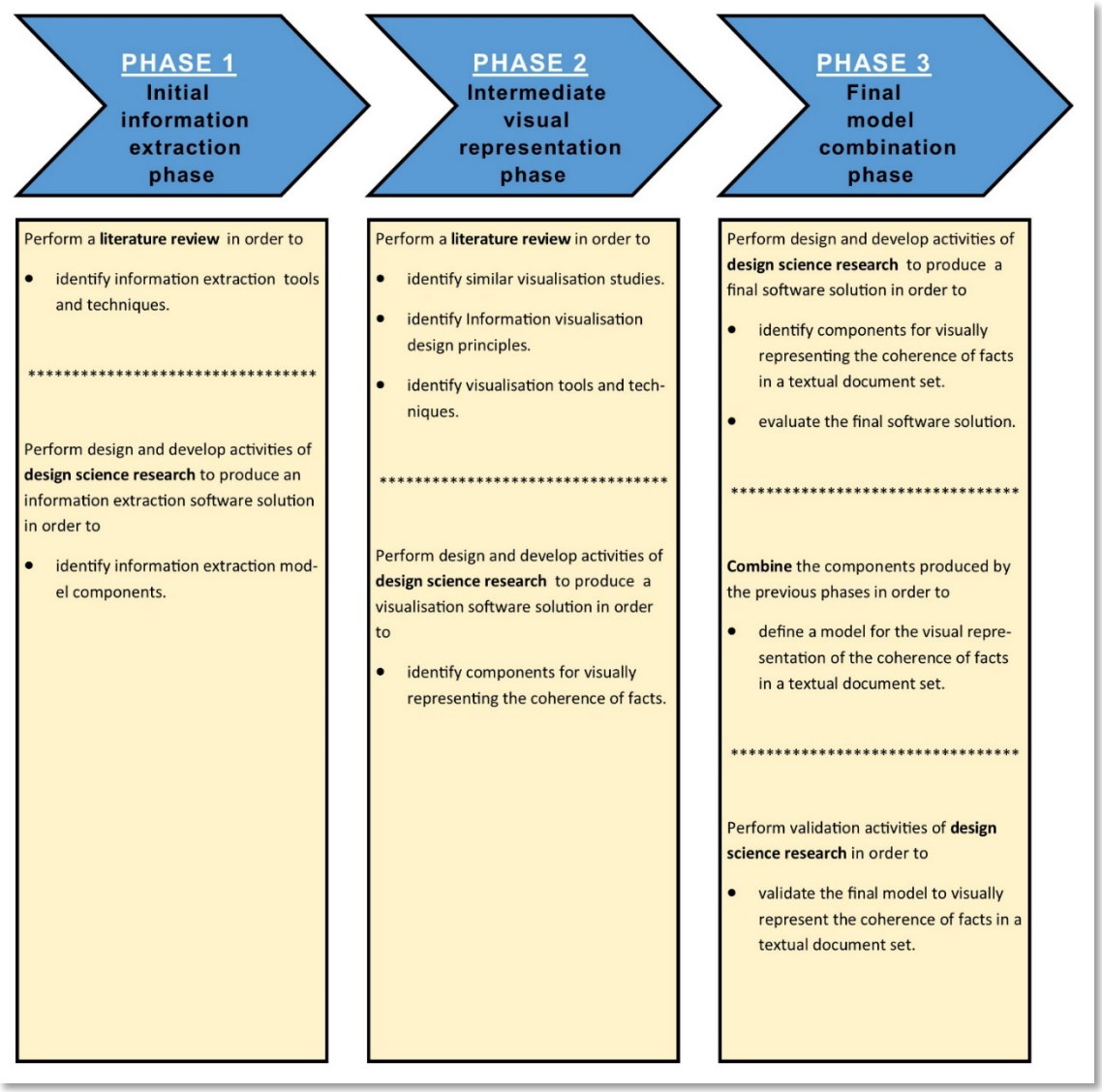


Figure 1.2: Research objectives and phases

Design science research, as a research methodology, is followed in order to develop a model for the visual representation of the coherence of facts in a textual record set. The design science research approach followed in this research study is discussed in detail in Chapter 2.

In order to identify visualisation design principles that could be applied to the design of an information visualisation interface, a literature review (see Section 4.1.3) is conducted during phase 2.

Having discussed the problem statement, research questions and research objectives of the research study, the next section briefly outlines the research design and methodology followed to achieve the objectives of the research.

1.3 RESEARCH DESIGN AND METHODOLOGY

A pragmatic design science research approach to address the purpose of the study is followed. According to Gregor and Hevner (2013), the construction of a design science research artifact should be informed by descriptive *kernel theories*. For the purpose of the study several *kernel theories* were incorporated into the design of the final artifact including the coherence theory of truth (Candlish & Damnjanovic, 1989; Glanzberg, 2013; Roche, 2013; Young, 2013), Ackoff's hierarchy of human understanding (1989) and Shneiderman's visual information seeking mantra (1996). These theories and others are discussed in Section 2.2.1.1.

In order to inform the construction of the *final* model, a specific software solution consisting of two parts are created and instantiated, namely an *initial* information extraction and an *intermediate* visual representation part. As a result of the created software solution parts, model components for both the parts are identified. Consequently, the created parts inform the development of two models as follows:

- an *initial* model to extract information and transform unstructured, textual data into a structured form;
- an *intermediate* model for the visual representation of the coherence of information contained in a structured data form.

By combining the two specific software instantiations and resulting models, a *final* model for the visual representation of the coherence of information in a textual document set is defined.

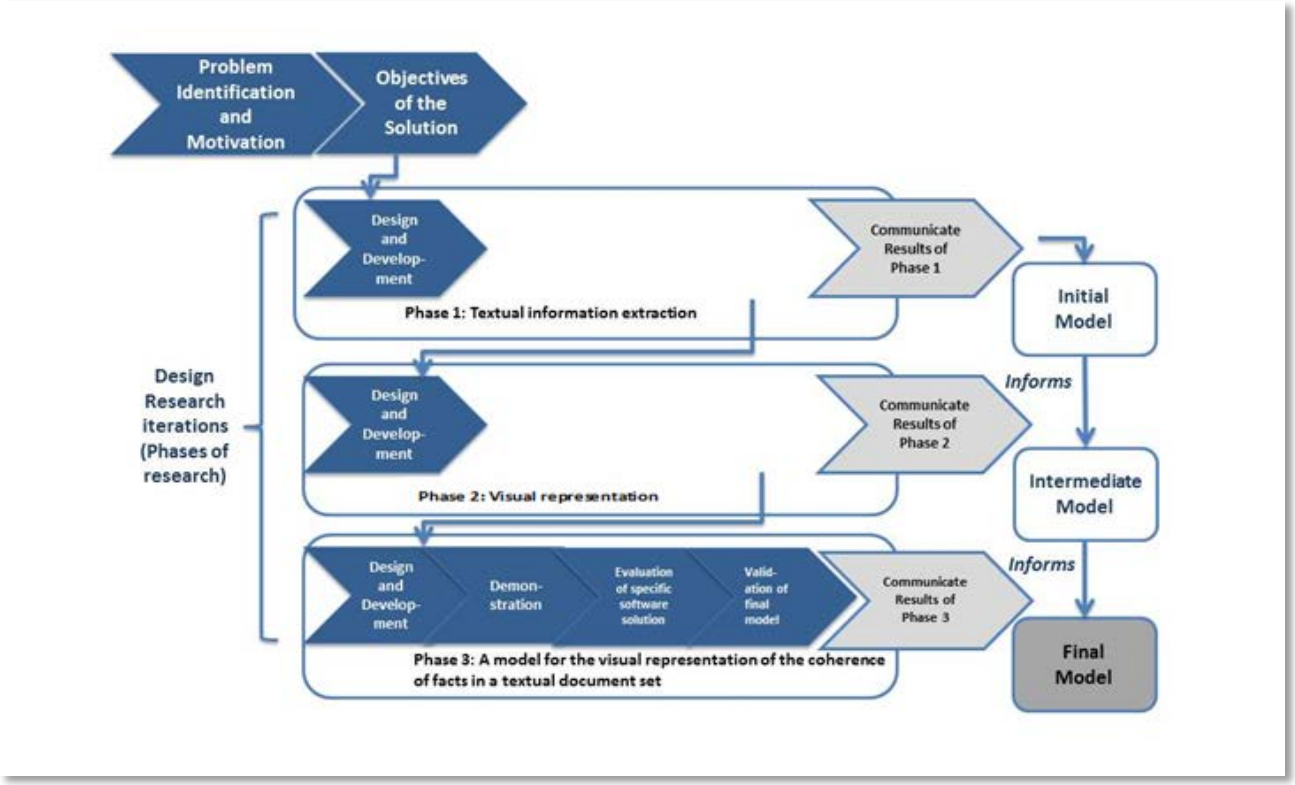


Figure 1.3: Adapted design science research process from Peffers *et al*. (2006)

The research is consequently conducted in three phases (depicted in Figure 1.3), adapting the design science research process model from Peffers *et al*. (2006), namely:

- *Phase 1: Information extraction*. An *initial* model is developed by starting off with a literature review, which is followed by the design, build and instantiation of the first part of the specific software solution. The process is described in Chapter 3.
- *Phase 2: Visual representation*. An *intermediate* model is developed by starting off with a literature review, which is followed by the design, build and instantiation of the second part of the specific software solution. The process in described in Chapter 4.
- *Phase 3: Final model.* A model for the visual representation of the coherence of facts in a textual document set is developed by combining the information extraction and visual representation instantiations and models from the preceding phases. The *final* model is described in Section 5.2.

The specific software solution instantiation is demonstrated to and evaluated by domain and Information Technology (IT) experts. The general model which emerged from the specific software solution is validated by IT experts.

The design science research process model of Peffers *et al*. (2006) is adapted to create the *initial* and *intermediate* models, ultimately informing the development of the *final* model. The process model consists of the following six processes for conducting design science research:

1. Identify the research problem and motivate the importance of solving the problem.
2. Define the objectives that the solution should achieve.
3. Design and develop an IT artifact to inform a solution for the problem.
4. Demonstrate the ability of the developed artifact to solve the problem.
5. Evaluate whether the built artifact provide a solution to the identified problem.
6. Communicate the results of the research.

The *final* software solution instantiation is evaluated and the resulting model for the visual representation of the coherence of facts in a textual document set is validated as follows:

- The *final* software solution instantiation is demonstrated to a number of participants (five experts from the domain described by the data and five IT experts) during interviews. Afterwards, the participants were asked to complete a feedback questionnaire in the form of a five-point Likert scale statement set. Pivot tables and charts for each of the statements presented in the feedback questionnaire are analysed in order to evaluate the final software solution instantiation.
- The *final* model is developed from the *initial* and *intermediate* models resulting from the corresponding design and build phases. The model is then demonstrated during interviews to the same IT experts to which the *final* software solution instantiation was previously demonstrated. The participants were asked to complete a feedback questionnaire in the form of a five-point Likert scale statement set. In order to evaluate the model, pivot tables and charts for each of the statements presented in the feedback questionnaire are analysed.

The statements presented to the participants during both the *final* software solution instantiation demonstration interviews and the *final* model presentation aimed to determine whether

- design principles where applied in the design of the visual representation,
- the final model provides components to extract facts from unstructured textual documents and
- whether the final model provides components to visually represent the coherence of facts in a document set.

Having briefly discussed the research methodology followed in the execution of the research, the next section discusses the importance of the research study.

1.4 IMPORTANCE OF THE RESEARCH STUDY

The research study has several significant benefits. In conducting the research study, it is envisaged that a better understanding of information extraction and information visualisation is acquired. It is envisaged that the better understanding will lead to an increase in the scientific knowledge base of both the research community and the researcher. The increase in knowledge of these two fields of study is also considered as having an increase in the number and quality of relevant IT artifacts built.

Grounded in the coherence theory, the research study demonstrates how information visualisation can be used to gain insight into the validity of domain facts contained in a textual document set. A contribution to the study of Indigenous Knowledge is made by presenting a proposed novel model for assisting the process of making an assumption on the validity of facts contained in Indigenous Knowledge, therefore contributing to the scientific evaluation and acknowledgement of Indigenous Knowledge.

The solution obtained can also be applied to other fields of studies where the records exhibit the same characteristics as defined in Section 1.2.1. The study can therefore be the starting point for similar research in a variety of other study fields such as Medical studies (e.g. patient health records), Criminology (e.g. crime reports) and Political Science studies (e.g. political speeches).

In addition the focus of this study is both on Information Extraction and Information Visualisation, therefore the result will contribute towards understanding the gap between the two research fields.

In the next section, the delineation and assumptions of the research study are set out.

1.5 DELINEATION AND ASSUMPTIONS

The delineation of the study and the assumptions made are set out as follows:

- The study only considers a set containing a finite set (a number of not more than 50 documents is proposed as appropriate) of original or transformed documents
- Although acknowledged as important, the research study does not take into account accessibility and usability issues with regard to the visual representation. These issues can be the subject of future studies.

- The research study does not attempt to find the most effective solution to visually representing the unstructured data, but rather a suitable option. Comparing and evaluating visualisation techniques are accordingly not considered within the scope of the research study. Subsequent research can be aimed at a more effective option by evaluating the effectiveness of the solution and adjusting the solution to increase its effectiveness.

- The research study does not attempt to find the most effective solution to extract domain facts from the textual source. Due to the scope and time of the master's degree study and the amount of research required to combine automatic information extraction methods with the visualisation research, the manual information extraction method is selected. Extracting domain facts from the source and finding a more effective information extraction solution can be the focus of future research.

- For the purpose of the study, it is assumed that if there exists a coherence in the facts contained in a set of textual records, then there possibly exists some truth in the contained facts. This proposition is based on the coherence theory of truth (Young, 2013).

- Some of the other classical theories of truth are the correspondence theory, Tarski's semantic theory, the pragmatic theory and the deflationary theory (Dowden & Swartz, 2013). Because the study of truth belongs to the field of Philosophy, it is beyond the scope of this study to compare and scrutinise existing theories of truth.

  However, the research study acknowledges that other theories of truth also exist. Candlish and Damnjanovic (1989) and Glanzberg (2013) can be referenced for more information on the subject of truth theories.

In the next section, the ethical considerations taken into account by the research study are briefly outlined.

1.6 ETHICAL CONSIDERATIONS

Ethical issues need to be considered to protect the integrity of the researcher as well as the integrity of the organisation, according to Burian, Rogerson and Maffei (2010). Burian *et al*. (2010) indicate that ethical considerations consist of three issues, namely to do no harm, to show respect and to remain just and fair.

For the purpose of the research study, the following ethical issues related to the use of the Huyser collection of Anglo-Boer War testimonies are considered:

- In doing no harm, the identities of persons have to be protected whenever the disclosure of their identities can cause harm. In order to protect the descendants of

individuals described in the testimonies, names are not used. Instead, original names are replaced with aliases.

- In showing respect, the use of archaic words contained in the records is carefully considered. Some of the words used in the testimonies are not in use anymore. Care is taken during the process of translating the testimonies into English to ensure that the essence of the meaning of the words was preserved.

- In remaining just and fair, the researcher's opinion should not influence the facts contained in the data set. The testimonies in the Huyser collection depict the events of the war from the viewpoint of civilian Boer people and therefore have an element of subjectivity. For the purpose of the study, to remain just and fair, the researcher does not reflect on personal opinions on the information contained in the testimonies.

Ethical clearance for conducting research was given by Unisa's College of Science, Engineering and Technology's (CSET) Research and Ethics Committee.

In the next section, the chapter outline of the dissertation is presented.

1.7 CHAPTER OUTLINE

The work conducted by the research study is provided in the next sections. The chapter outline of the dissertation is depicted in Figure 1.4 on the following page.

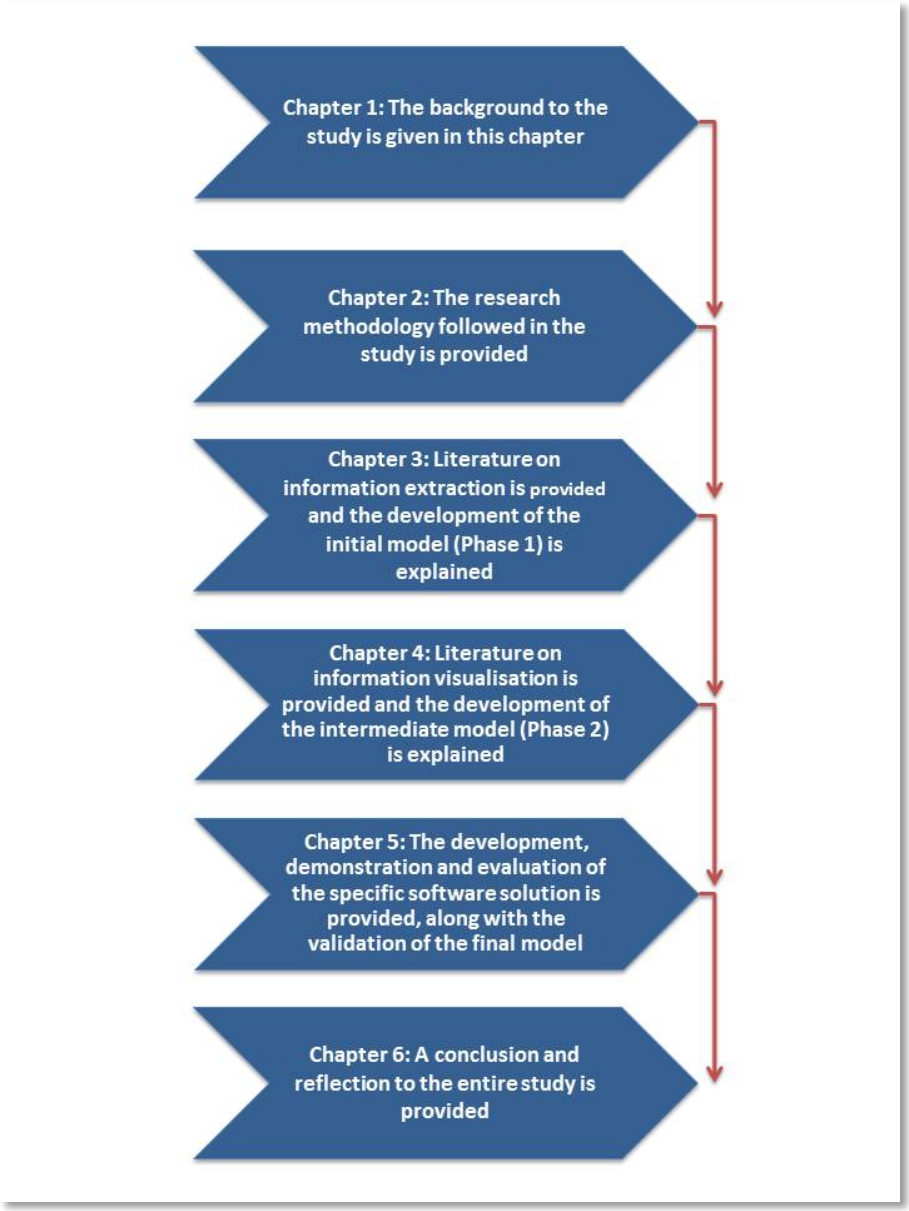


Figure 1.4: Chapter outline

After having provided background information, research purpose, research questions, research objectives, importance and chapter outline of the research study, the research design and methodology used to achieve the purpose of the study is discussed in the next chapter.

CHAPTER 2: RESEARCH APPROACH AND METHODOLOGY

The previous chapter provided an introduction to the research which consisted of the research background, research purpose and questions, summary of the research design and methodology, importance of the research, the delineations and assumptions of the research and the ethical consideration.

The purpose of the research, together with the research approach, is considered by Hofstee (2009) as vital to the success of any research study. The purpose of the study, namely to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set, was discussed in Chapter 1. In this chapter, the research approach to achieving the purpose of the study, is outlined in more detail.

According to Creswell (2009, 2014), research approaches are plans and procedures for conducting research. The research plan involves an overall decision about the approach taken to achieve the research goals. The plan is informed by the philosophical assumption on which the study is built (called research paradigm), the research strategy (called research methodology) and the research method (Creswell, 2009, 2014).

Firstly, the selected research paradigm is considered; followed by a discussion of the selected research methodology. The research method followed to achieving the purpose of the study is then discussed next. The chapter is concluded by a consideration of the reliability and validity of the research.

2.1 RESEARCH PARADIGM

Research paradigms are regarded by Wahyuni (2010) as fundamental beliefs and assumptions that effects the way a researcher conducts research, which includes the choice of a particular research methodology. Creswell (2014) uses the term *worldview* to indicate that a research paradigm describes how the researcher sees the world from a general philosophical view, influencing the how the research is conducted.

According to Hussain, Elyas and Nasseef (2013) and Wahyuni (2010) existing research paradigms can be distinguished by considering the *ontology* and *epistemology* of each. The *ontology* of a study describes how reality is perceived (Scotland, 2012; Wahyuni, 2010) and what constitutes reality (Scotland, 2012). *Epistemology* is concerned with the nature and forms of knowledge (Scotland, 2012), describing the research approach to generating, understanding and using knowledge (Wahyuni, 2010). Krauss (2005) argues that the *epistemology, ontology* and *methodology* of a study are interrelated in that *ontology* involves the philosophy of reality, *epistemology* is engaged with how one gets to know that reality and *methodology* defines the practices to acquire that knowledge.

Having described what a research paradigm is and how a paradigm is distinguished from other paradigms, major paradigms considered by Creswell (2014) and Hussain *et al*. (2013) are presented:

- *Positivism*. The positivist researcher believe that knowledge can acquired by making observations and measurements, producing models of how the world works (Oates, 2006). Positivist researchers regards reality as consisting of what is available to the human senses (Gray, 2013). Creswell (2014) states that in addition, the post-positivist paradigm considers the truth after positivism. Post-positivist researchers challenge the idea of the absolute truth about knowledge and assumes that we cannot be positive about knowledge claims when studying the behaviour and actions of people (Wahyuni, 2010).

- *Interpretivism*. Also known as *Constructivism* (Wahyuni, 2010). The constructivist or interpretivist researcher believe that understanding and knowledge is constructed from interpretation, assuming that there are no facts, only interpretations (Hussain *et al*., 2013). “Reality is constructed by social actors and people’s perception of it” (Wahyuni, 2010, p. 71).

- *Critical theory*. Also described as the *Transformativ*e theory (Mackenzie & Knipe, 2006). According to Oates (2006, p. 296), critical research “is concerned with identifying power relations, conflicts and contradictions, and empowering people to eliminate them as sources of alienation and domination”. Current held values and assumptions are questioned and conventional social structures are challenged by critical researchers (Gray, 2013), driven by a desire for transformative change (Hussain *et al*., 2013). Reality is considered as something that is tangible and composed of historical structures and knowledge is derived from social context (Hussain *et al*., 2013).

- *Pragmatism*. The pragmatic researcher starts with an emphasis on the research problem, using several approaches to understand reality (Creswell, 2014; Wahyuni, 2010). Truth is based on what works at the time to address the research problem (Creswell, 2014; Gray, 2013), acquiring knowledge from actions, situations and consequences (Creswell, 2014). In order to address the research problem at hand the researcher can apply multiple systems of philosophy, with a freedom to choose the research methods, techniques and procedures (Creswell, 2014).

The *epistemology* and *ontology* of each of these major paradigms as provided by Wahyuni (2010) are listed in Table 2.1 on the following page.

Table 2.1: *Epistemology* and *ontology* of *Positivism*, *Interpretivism*, *Critical research* and *Pragmatism* as provided by Wahyumi (2010)

Paradigm	Epistemology	Ontology
Positivism	Only observable phenomena can provide credible data, facts. Focus on causality and law-like generalisations, reducing phenomena to simplest elements.	External, objective and independent of social actors.
Interpretivism	Subjective meanings and social phenomena. Focus upon the details of situation, the reality behind these details, subjective meanings and motivating actions.	Socially constructed, subjective, may change, multiple.
Critical theory	Only observable phenomena can provide credible data, facts. Focus on explaining within a context or contexts.	Objective. Exist independently of human thoughts and beliefs or knowledge of their existence, but is interpreted through social conditioning (critical realist).
Pragmatism	Either or both observable phenomena and subjective meanings can provide acceptable knowledge dependent upon the research question. Focus on practical applied research, integrating different perspectives to help interpret the data.	External, multiple, view chosen to best achieve an answer to the research question.

In order achieve the aim of the study, the researcher chose to follow a pragmatic approach. This decision is guided by the researcher's own work experience of solving problems in a practical manner, believing that truth can be based on a workable, practical solution.

Having discussed the philosophical approach to research adopted by the researcher, the next section gives a discussion of the overall research methodology followed to conduct the research study.

2.2 RESEARCH METHODOLOGY

In the previous section, the research paradigm or philosophical approach to research adopted by the researcher was discussed. According to Hussain *et al*. (2013) and Gray (2013), the particular methodology chosen by the research to conduct research, is influenced by the philosophical underpinning or research paradigm adopted by the researcher. In the particular case of this research study the decision to choose a pragmatic approach, guided the choice of the particular methodology.

Hussain *et al*. (2013) states that the research methodology refers to general principles which describe how the world is investigated and how it is demonstrated that the knowledge generated from the investigation is valid. Scotland (2012) describes the methodology as a strategy or plan of action "which is concerned with why, what, from where, when and how data is collected and analysed". In addition, Oates (2006) describes the strategy or

methodology as the overall approach to answering the research question; listing and describing the following possibilities:

- *Survey*. Data is collected from a large group of people in a systematic manner. The data is then analysed to look for patterns. The results are then generalised to a larger population than the group from which the data was collected.

- *Experiment*. Hypotheses are tested by investigating cause and effect relationships. The goal is to prove or disprove a causal link between a factor and an observed outcome.

- *Case study*. An instance of a phenomena is investigated in order to obtain a rich insight into the life of that case and its complex relationships and processes.

- *Ethnography*. The researcher spend time in the field, taking part in the life of people. The goal is to gain an understanding of the culture and customs of a particular group of people.

- *Action research*. Research is conducted by planning to do something in a real-world situation, doing it and then reflecting on what has happened or what was learned.

- *Design and creation*. The research focus on the creation of a new IT artifact which can be a new computer based application, construct, model or method. Johannesson, Perjons and Bider (2013) state that design and creation is also known as design science research. For the remainder of the dissertation, the term design science research will be used.

In order to achieve the purpose of the study, an overall design science research approach is followed for the following reasons:

- The purpose of the research is to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set. A model is considered by various authors (Carlsson, 2006; Cole, Purao, Rossi, & Sein, 2005; Hevner & Chatterjee, 2004; March & Smith, 1995) as an artifact. Furthermore, design science research is a problem-solving paradigm that attempts to create and evaluate new innovations in the form of IT artifacts (Hevner & Chatterjee, 2004; Niehaves, 2007). Design science research is consequently considered to be an appropriate research approach for this study since both design science research and this study have an artifact (a model for the visual representation of the coherence of facts in a textual document set) as end-goal.

- Hevner and Chatterjee (2010) state that design science research is an appropriate approach to solve “wicked” problems, namely problems with the following characteristics:

    - The requirement and constraints of the problem are unstable.

- The interactions between components of the problem are complex.
- The problem exhibits an inherent flexibility to change the design process and design artifact.
- The problem depends on human cognitive abilities to produce the solution.
- The problem depends on human social abilities to produce the solution.

By considering these characteristics of “wicked” problems, the problem of extracting information from textual records and the visualisation of the coherence of facts in the extracted records are considered to be “wicked” for the following reasons:

- The requirement and constraints of the problem are unstable in that they originate in the need for the research. The requirements and constraints are defined as part of the research and are not known before the research.
- The interactions between components of the problem are complex in the sense that the interactions are not known at the start of the research and therefore only become known as a result of the study.
- The design of the processes and the final model come into being as a result of the research. Only during the research are the processes, specific software solutions and models defined and built. As a result, there is an inherent flexibility to make changes to the processes and artifacts.
- In order to produce the solution, an in-depth understanding of the purpose of the research as well as an amount of creativity to design the visual interface is required.
- Human social abilities are required in order to demonstrate the specific final software solution and model to a participant audience. An amount of interaction between the researcher and the audience is required to explain the problem and to demonstrate the solution.

As a result of both design science research and this study having an artifact as end-goal, and because the problem identified by the research is considered to be “wicked” according to the preceding requirements, design science research is selected as research approach for conducting the research study.

Design science research as a methodology for conducting the research study is discussed in the next section.

2.2.1 Design science research

Cole *et al*. (2005) postulate that Information Systems (IS) research should solve practical problems and, in addition, contribute to the existing knowledge base of science.

Furthermore, according to Hevner, March, Park and Ram (2004), Carlsson (2006), Niehaves (2007) and Cole *et al*. (2005), IS research is mostly conducted either through a behavioural science approach or a design science approach.

Carlsson (2006) considers the behavioural science approach as the mainstream approach to IS research and postulates that behavioural science is description and explanatory driven whereas design science is prescriptive driven. In addition, Niehaves (2007) considers behavioural science and design science as complementary parts of IS research. Research conducted through a behavioural science approach attempts to develop and justify theories that explain the analysis, design, implementation, management and use of information systems, whereas design science research is a problem-solving paradigm that attempts to create and evaluate new innovations in the form of IT artifacts (Cole *et al*., 2005; Hevner *et al*., 2004; Niehaves, 2007).

Design science research is discussed in the following sections by examining its philosophical views (Section 2.2.1.1), artifacts produced (Section 2.2.1.2) and knowledge contributions made (Section 2.2.1.3). In addition, guidelines (Section 2.2.1.4) for conducting design science research as well as its research cycles (Section 2.2.1.5) are discussed. The section on the design science research is concluded by considering the advantages and disadvantages (Section 2.2.1.6) of the methodology.

2.2.1.1 Philosophical views of design science research

Philosophically, design research is considered by various authors (Cole *et al*., 2005; Hevner *et al*., 2004) to be pragmatic. Carlsson (2006), on the other hand, suggests that although many design science research publications does not explicitly discuss research *ontology* and *philosophy* issues, most design science research seems to be based on *positivism*, *traditional realism* or *pragmatism*. In conducting this research study, a *pragmatic* research approach is followed as indicated in Section 2.1.

Furthermore, Cole *et al*. (2005) state that the *ontology* of design science research assumes that the phenomenon of interest changes through the application of the research process and they further state that, with regard to this phenomenon, an artifact realises as a result of the research. The *epistemology* of design science research, according to Cole *et al*. (2005), resides in knowledge being obtained by means of intervention to cause change and reflection on the change.

Gregor and Hevner (2010, 2013) state that there are two schools of thought in the design science research community. The first school which they termed as the *design-theory* camp, focus on the importance of grounding the design of the research artifact on a design theory. The second school which they termed as the *pragmatic-design* camp, is considered to be “agnostic” to the need for a design theory and place more importance on the artifact itself. The different schools in addition place different emphasis on the contribution made by design science research. According to Gregor and Hevner (2013), the harmonisation of the two views will improve the rigour of design science research. For the purpose of the research

study, these two views are both considered by focusing on both the utility and value of the designed artifacts, as well as grounding the design of artifacts in *kernel theories*. As a result of the study, a theory in the form of a model for the visual representation of the coherence of facts in a textual record set is proposed. In addition, the research study propose that a visual representation of the coherence of facts in a set can assist with the process of making a decision about the set's truth value.

In order to impact the rigour of design science research, Gregor and Hevner (2013) propose that the construction of the research artifact should be informed by descriptive *kernel theories*. These theories are used to justify the design of the artifact, which include any theory from the IT domain as well as any external domain (Grergor and Hevner, 2010). The overall design of the model for the visual representation of the coherence of facts in a textual record set is based on several *kernel theories*:

- The coherence theory of truth (Candlish & Damnjanovic, 1989; Glanzberg, 2013; Roche, 2013; Young, 2013). Grounded on this theory, the study assumes that the coherence of a fact in a textual document set is an indication of the possible truth value of the fact.

- Ackoff's hierarchy of human understanding (1989) which considers *Knowledge* to be derived from *Information*, which in turn is derived from *Data*. For the purpose of the study, the set of Boer War testimonies is considered as *Data*. The extracted facts is considered to be *Information* and by visually representing the coherence of the facts, it is assumed that the user gains some *Knowledge* about the truth value of the facts.

- Design principles and guidelines compiled by means of a literature review which incorporated, amongst others, the following:

    - the Eppler and Burkhard Knowledge Visualisation framework (2007);

    - Forsell and Johansson's heuristic set for Information Visualisation evaluation (2010);

    - Elmqvist and Fekete's guidelines for implementing visual hierarchical aggregates (2010);

    - Shneiderman's visual information seeking mantra (1996), which provides seven typical tasks that a user performs when searching for information.

Having discussed the philosophy of design science research and the use of *kernel theories*, the artifacts produced by a design science research study is discussed in the next section.

2.2.1.2 Artifacts produced by design science research

According to March and Smith (1995) the artifacts produced by a design science research study can be *constructs*, *models*, *methods* or *instantiations*. The statement is supported by

various authors *(Carlsson, 2006; Cole et al*., 2005; Hevner *et al*., 2004; Iivari, 2007). Table 2.2 lists short definitions of the types of artifacts as provided by March and Smith (1995).

Table 2.2: Design science research artifacts

Type of artifact	Definition
Construct	Forms the vocabulary of a domain and is used to describe domain problems and to specify the solutions to the problems (March & Smith, 1995).
Model	A set of statements that express relationships among constructs. A "representation of how things are" (March & Smith, 1995, p. 256).
Method	A set of steps that can be used to perform a task, including algorithms and guidelines (March & Smith, 1995).
Instantiation	"The realization of an artifact in its environment" (March & Smith, 1995, p. 258).

Gregor and Hevner (2010) suggest that, in addition, to *constructs*, *models*, *methods* and *instantiations*, any designed solution that solves a problem in context should also be added to the list of possible artifacts produced by design science research.

For the purpose of achieving the objectives of the research study, the following artifacts are produced by the research study:

- An *initial* solution instantiation for extraction facts from a set of textual records. The end-product of the specific *initial* solution is a database schema populated with domain facts contained in a set of textual documents.

- An *initial* model to extract facts and transform unstructured textual data into a structured form is informed by the instantiation of the *initial* solution instantiation.

- An *intermediate* software solution instantiation for visually representing the coherence of facts in a document set. The end-product of the specific intermediate software solution is a visual representation of the coherence of facts contained in a document set.

- An *intermediate* model to visually represent the coherence of facts contained in a structured data form is informed by the instantiation of the *intermediate* software solution instantiation.

- A *final* specific solution instantiation providing a method for extracting domain facts from a textual record set as well as a visual representation of the coherence of facts in the set. The *final* specific solution is constructed by combining the *initial* information extraction solution as well as the *intermediate* visual representation software solution.

- A *final* model for the visual representation of the coherence of facts in a textual document set is informed by the instantiation of the *final* specific solution. The model

is consequently a conglomeration of the *initial* information extraction and the *intermediate* visual representation models.

Having described the design science research artifacts produced by the research study, the next section articulates the knowledge contributions made these artifacts.

2.2.1.3 Knowledge contributions made by design science research

Hevner *et al*. (2004) postulate that the difference between routine artifact building and design research is in the contribution made to the existing knowledge base of science. The contribution that can be made by a design science research project is considered by Gregor and Hevner (2010, 2013) to be on one of three levels (from less abstract and more specific to more abstract and more general) as follows:

- **Level 1:** contributions which include products and processes;
- **Level 2:** contributions which include design principles such as models and methods;
- **Level 3:** contributions which include emergent design theories about the phenomena being studied.

Gregor and Hevner (2010) also suggest that a novel IT product can also be considered to be a research contribution. The research study contributes on both Level 1 and Level 2 of the Gregor and Hevner (2010, 2013) research contribution levels:

- **Level 1:** a product in the form of a specific software solution instantiation to visually represent the coherence of facts in a textual document set. The specific software solution is comprised of an *initial* information extraction and an *intermediate* visual representation solution. In addition, the *final* software solution instantiation itself is considered to be a research contribution because it presents a novel mechanism for studying domain information contained in textual data sources.
- **Level 2:** a model for the visual representation of the coherence of facts in a textual document set. The model is comprised of an *initial* information extraction and an *intermediate* visual representation model.

In addition, Gregor and Hevner (2013) provide a framework for categorising the possible contribution of any design science research project. The framework, as depicted in Figure 2.1 on the following page, categorises the contribution made by any design science research project according to two dimensions, which are based on the existing knowledge about the problem and the maturity of the solution domain. The contribution made by the research study is indicated on the figure.

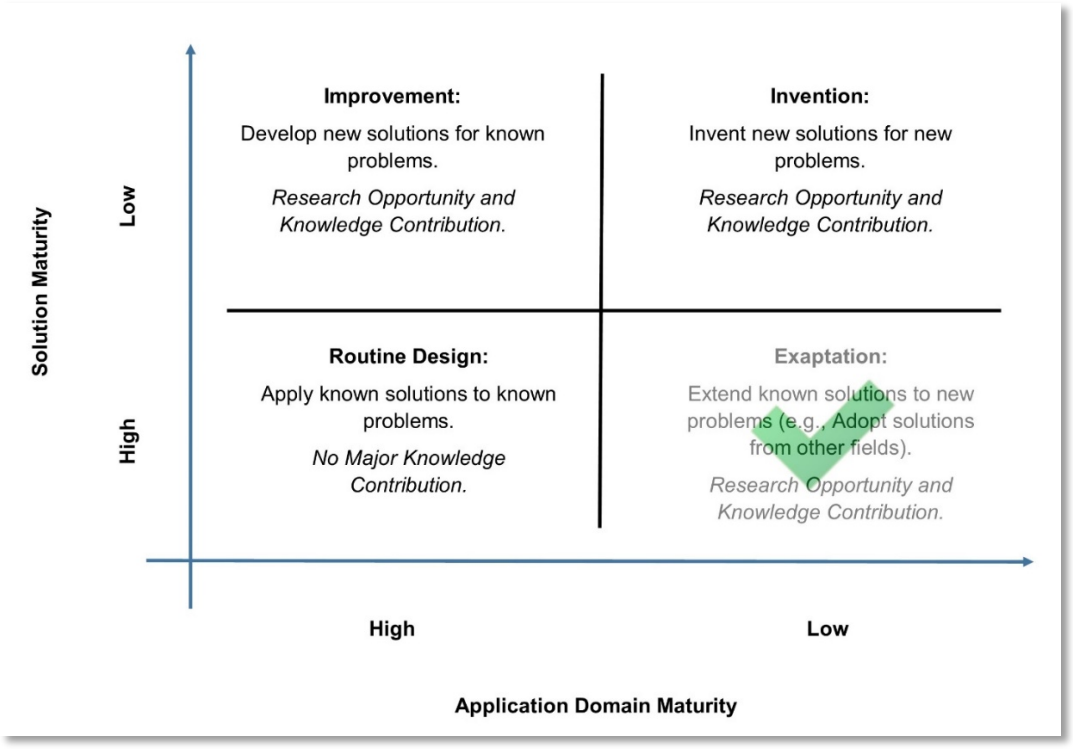


Figure 2.1: DSR knowledge contribution framework (Gregor & Hevner, 2013)

The *x-axis* represents the maturity of the problem (i.e. the degree to which the problem is known) and the *y-axis* represents the maturity of possible solutions (i.e. the degree to which current solutions exist for the problem). Therefore, Gregor and Hevner (2013) suggest that the contribution made by a design science research study can be a *routine design*, an *improvement*, an *invention* or an *exaptation*. For the purpose of the research study, a contribution to knowledge is made in the form of an *exaptation*:

> a) known solutions were extended (visualisation design principles and guidelines from the literature were applied to the design and the *zoomable circle-packing* technique provided by Bostock (2012) was used as starting point for the visual representation) for
>
> b) a new problem (visually representing the coherence of facts in a textual document set in order to make an assumption on the validity of the facts).

Iivari (2007) explains that three knowledge types are produced by IS research, namely *conceptual knowledge* (concepts, constructs, classifications and conceptual frameworks), *descriptive knowledge* (observational facts, theories and hypotheses) and *prescriptive knowledge* (design product and design process knowledge). Gregor and Hevner (2013), on the other hand, only considers *descriptive* (what knowledge) and *prescriptive* knowledge (how knowledge) as knowledge types produced by design science research.

The research study produce *prescriptive knowledge* in the form of an Information Visualisation design product, as well as a model that abstracts the process to extract and transform facts contained in textual documents into a structured format and to visually represent the coherence of the facts. The model is a conglomeration of an *initial* information extraction model and an *intermediate* visual representation model.

Having described the possible contributions that a design science research study can make, the next section considers Hevner *et al*.'s (2004) guidelines for conducting design science research.

2.2.1.4 Guidelines for conducting design science research

In executing the research, the study subscribed to Hevner *et al*.'s (2004) seven guidelines for conducting design science research. The seven research guidelines are listed in Table 2.3.

Table 2.3: Design science research guidelines (Hevner *et al*., 2004)

Guideline	Short description	Application to this research study
1. Design is an artifact.	The research should construct an artifact which can be either a construct, a model, a method or an instantiation.	The research produced a specific software solution instantiation which is comprised of an *initial* information extraction solution and an *intermediate* visual representation solution. The instantiation of the two solutions informed the development of an *initial* information extraction model and an *intermediate* visual representation model. Combining the two models resulted in a *final* model for the visual representation of the coherence of facts in a textual document set.
2. Problem relevance.	The objective of the research should be to develop an IT-based solution to a relevant problem.	To identify the components that can be combined into a model to visually represent the coherence of facts in a textual document set. It is argued that by visually representing the coherence of facts contained in a document set, the user can make an assumption about the validity of the facts and the document set.
3. Design evaluation.	The utility of the produced artifact should be evaluated.	An instantiation of *final* software solution was demonstrated to five Anglo-Boer War domain experts as well as to five technical IT experts during interview sessions. The audience was required to complete a feedback questionnaire in the form of a five-point Likert scale statement set. Pivot tables and charts for each of the statements presented to the participants were then analysed in order to evaluate the artifact. The *final model* was presented to the same IT experts to which the *final* software solution was previously demonstrated. They were also required to complete a feedback questionnaire in the form of a five-point Likert scale statement set. Pivot tables and charts for each of the statements presented to the participants were then analysed in order to evaluate the *final* model.

Guideline	Short description	Application to this research study
4.Research contributions.	The research should contribute to the existing knowledge base of science.	It is proposed that the research leads to a better understanding of Information Extraction and Information Visualisation. Furthermore, a model for visually representing the coherence of facts in a textual document set is developed, thereby providing a model for making an assumption on the validity of facts contained in records describing facts of a domain, for instance, Indigenous Knowledge. The contributions made by the study are discussed in more details in Section 1.4, Section 2.2.1.3 and Section 6.3.
5. Research rigour.	Rigorous research methods should be applied to the construction and evaluation of the artifact.	Rigour is ensured by following a referenced research method (design science research process model of Peffers *et al*. (2006)), by informing the design of the *initial* information extraction and *intermediate* visual representation instantiations and models from an existing science knowledge base and by adding the research results and its evaluation to the existing science knowledge base.
6. Design is a search process.	Design science research requires a search process to find an effective solution.	In the process of finding a solution, current information extraction techniques, information visualisation techniques, software tools and software libraries were visited in order to find a solution. The search process involved a review of literature presented in Section 3.1 and Section 4.1. Feedback received during the demonstration of the resulting artifact will be used at a later stage (out of scope of the research study) to reiterate the design of the solution.
7. Communication of research.	The results of the research should be effectively communicated.	The results of the research are communicated through this dissertation and through conference papers (Engelbrecht *et al*., 2014, 2016; Engelbrecht & Botha, 2015) and a journal article (Engelbrecht *et al*., 2015).

Having outlined design science as chosen research paradigm for conducting the research study, the research cycles of design science research are discussed briefly.

2.2.1.5 Design science research cycles

The design science research process to produce the knowledge contribution of the research, consists of four research cycles as defined by Drechsler and Hevner (2016). The four design science research cycles as depicted in Figure 2.2 on the following page consist of a *change and impact cycle*, a *relevance cycle*, a *design cycle* and a *rigour cycle*. The *change and impact* cycle was only recently added by Drechsler and Hevner (2016). Previous literature only defined three cycles as part of design science research.

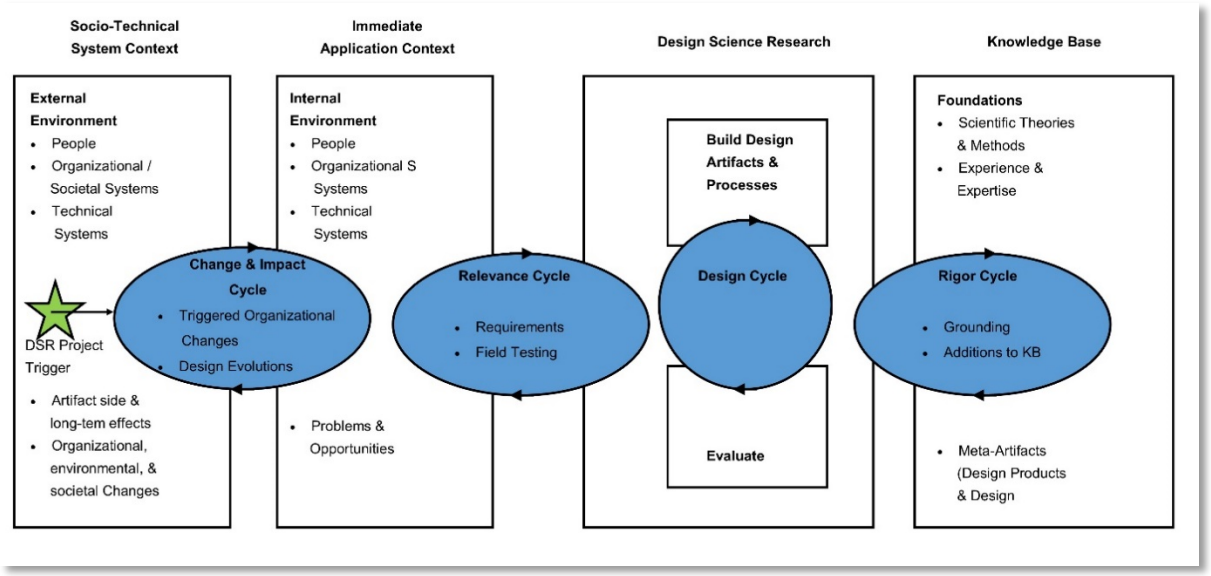


Figure 2.2: Design science research cycles (Drechsler and Hevner, 2016)

Change and impact cycle

During the *change and impact* cycle, the impact of the research on the external environment is considered (Drechsler and Hevner, 2016). The previous three-cycle proposal of Hevner (2007) considered the relevance of the research on a broad level. According to Drechsler and Hevner (2016), the additional cycle in particular considers the impact of the research on a context wider than just the immediate context of the application. Although indicated as the project trigger for a design science project, they state that “we do not consider iterations through the CI cycle as necessarily mandatory for DSR” (Drechsler and Hevner ,2016, p. 5).

For the purpose of the research study the impact on fields other the originating field of than Indigenous Knowledge are also considered. For instance it is proposed that the final solution and model can be applied to various other fields of study including History and Law. In particular the domain of History is indicated as an external impact area by using Anglo-Boer War testimonies for developing the model.

Relevance cycle

During the *relevance cycle,* input is provided to the research by the contextual environment. The input includes the requirements for the research, as well as the criteria for evaluating the research results (Hevner, 2007).

The NIKMAS research project is considered to be the contextual environment providing input into the study. The development of the NIKMAS system to capture and protect Indigenous Knowledge highlighted the phenomena that certain facts contained in

Indigenous Knowledge descriptions were in coherence. Based on the theory that “the truth of any (true) proposition consists in its coherence with some specified set of propositions” (Young, 2013, p. 1), it soon became apparent that providing a visual representation of the coherence of facts in a document set could assist with the process of making a decision about the validity of the document set. Therefore, the main requirement of the research study was to provide a model for the visual representation of the coherence of facts in a textual document set.

Design cycle

The *design cycle* consists of the activities to build and evaluate the design science artifacts (Hevner, 2007). For the purpose of the research, an *initial* information extraction solution, an *intermediate* visualisation representation solution and a *final* information visualisation solution was designed, built and instantiated.

After each solution was built using various tools and techniques as described in Chapters 3 and 4, the *final* information visualisation solution instantiation (a conglomeration of the *initial* and *intermediate* solutions) was demonstrated to a number of IT experts and Anglo-Boer War experts (experts in the domain described by facts in the document set used as source). The demonstrations were in the form of interviews. After each interview, each participant was required to complete a feedback questionnaire in the form of a five-point Likert scale statement set. Each statement was purposely defined in order to evaluate whether design principles and guidelines were considered in the design of the visual representation; whether the *final* information visualisation solution instantiation consisted of components to extract facts and to transform unstructured textual data into a structured form; and whether the *final* solution visually represented the coherence of facts contained in a textual document set.

After the instantiation of the *initial* and *intermediate* solutions, models for each were developed. The instantiation of each individual solution part was used to inform to development of the corresponding models. The resulting *final* model was demonstrated to the same IT experts to which the *final* solution was demonstrated. The experts were asked to complete a feedback questionnaire in the form of a five-point Likert scale statement set. In order to evaluate the model, pivot tables and charts for each of the statements presented in the feedback questionnaire were analysed. The analysis of the statements aimed to determine whether the *final* model provided components to extract facts from unstructured textual documents and whether the *final* model provides components to visually represent the coherence of facts in a textual document set.

Rigour cycle

The *rigour cycle* provides a knowledge base of scientific theories, methods, experiences and expertise in the research domain. The knowledge is used to inform the design of the artifacts as required by the research. According to Iivari (2007), it is the rigour of constructing IT artifacts during science research that distinguishes the research from merely IT artifact development. The *design cycle* of the research was informed by the existing knowledge in the domains of Visualisation and Information Extraction.

In order to get a better understanding of *Visualisation*, various views on visualisation concepts had to be considered. The result was working definitions of *Visualisation*, *Knowledge Visualisation*, *Information Visualisation* and *Document Visualisation*. With these working definitions, the researcher concluded that the subject of the research is Information Visualisation - the transformation from lower-level data to visual representation of meanings extracted from the data, where the extraction is either a computational process or a human transcription process; the aim is to explore data and create new insights.

After identifying Information Visualisation as the subject of the study, a set of Information Visualisation design principles and guidelines were constructed in Section 4.1.3 by means of a literature review. In addition to the acquisition of relevant information visualisation background knowledge, various information extraction tools and techniques were identified by means of a literature study. The decision to employ a supervised information extraction method was informed by the knowledge base of expert peer opinions during the presentation of the intended research at the SAICSIT 2014 Master's and Doctoral Symposium (SAICSIT 2014 Master's and Doctoral Symposium, 2014) and at the FASTAR/Espresso Workshop 2014 (Engelbrecht, 2014).

Grergor and Hevner (2013) propose that the construction of the research artifact should be informed by descriptive *kernel theories* in order to improve the rigour of the study. The overall design of the final model for the visual representation of the coherence of facts in a textual record set is based on several *kernel theories* including the coherence theory of truth (Candlish & Damnjanovic, 1989; Glanzberg, 2013; Roche, 2013; Young, 2013); Ackoff's hierarchy of human understanding (1989); the Eppler and Burkhard knowledge visualisation framework (2007); Forsell and Johansson's heuristic set for information visualisation evaluation (2010); Elmqvist and Fekete's (2010) guidelines for implementing visual hierarchical aggregates (2010); and Shneiderman's visual information seeking mantra (1996). Section 2.2.1.1 can be referenced for more information on these *kernel theories*.

Rigour, however, does not only require that the existing knowledge base should inform the design research, but it also requires that the result of the research should be added to the existing knowledge base (Hevner, 2007). In addition to the final dissertation, results of the research were added to the scientific knowledge base by means of conference papers (Engelbrecht *et al*., 2014, 2016; Engelbrecht & Botha, 2015) and a journal article (Engelbrecht *et al*., 2015).

Having discussed the cycles of design science research in this section, the advantages and disadvantages of design science research is considered next.

2.2.1.6 Advantages and disadvantages of design science research

Oates (2006) lists the advantages of design science research, which he defines as the *design and creation research strategy*. These advantages are the following:

- The output of the research is something which is tangible.
- It is suitable for people with a technical background.
- It is the usual approach in the field of computer science.
- There are many opportunities to make a contribution to the body of knowledge through the building of artifacts.

The disadvantages of the *design and creation research strategy,* according to Oates (2006), are the following:

- The researcher may have to justify why it is research and why it is not just the building of an artifact.
- Technical and artistic skills are required.
- Generalisation from a single artifact might be difficult to achieve.
- The lifespan of the artifact might depend on the presence of the researcher.
- Rapid advances in technology might invalidate the research results before they have been published.

The research study will resolve the disadvantages of the *design and creation research strategy* as defined by Oates (2006) as follows:

- The research will apply a proven research process that is applicable to the design science research paradigm. The design science research process model as defined by Peffers *et al*. (2006) and proposed by Hevner and Chatterjee (2010) as a suitable process model for conducting design science research, is followed for conducting the research study.
- The requirement for technical and artistic skill for conducting the research is resolved by the researcher's professional software development skill level.
- The aim of this research study is not to generalise from a single artifact, but rather to provide a possible solution model for visually representing the coherence of facts

contained in a textual document set. Therefore, the problem of generalisation from a single artifact is not applicable to this research study.

- In order to solve the problem of the lifespan of the artifact being dependent on the lifespan of the researcher, the application code is shared with colleagues.

- In order to solve the problem of rapid advances in technology that might invalidate the research results before they have been published, it is important that the researcher keep up to date with recent advances and also publish interim results as soon as possible. Interim results have been published in conference papers (Engelbrecht *et al*., 2014, 2016; Engelbrecht & Botha, 2015) and a journal article (Engelbrecht *et al*., 2015).

Having discussed the design science research paradigm, the next section discusses the specific research method followed in order to conduct the design science research of the study.

2.3 RESEARCH METHOD

The specific method followed in order to conduct the design science research of the study is articulated next in this section. In order to achieve the purpose of the study, two specific solution instantiations, namely an *initial* information extraction solution and an *intermediate* visual representation solution are created. The combination of both resulted in a *final* information visualisation solution. These specific instantiations respectively inform the development of an *initial* information extraction model and an *intermediate* visual representation model. By combining the two models, a resulting *final* information visualisation model is constructed. It is therefore stated that the instantiations of the *initial* information extraction solution and the *intermediate* visual representation solution informed the development of a model for visually representing the coherence of facts in a textual document set – achieving the purpose of the research study.

According to Gregor and Hevner (2010), there are a number of methods available to conduct design science research, which include methods provided by Nunamaker, Chen and Purdin (1990), Purao (2002) and Peffers *et al*. (2006). Hevner and Chatterjee (2010) propose the use of the design science research process model as defined by Peffers *et al*. (2006) for conducting design science research. According to Peffers *et al*. (2006), the model provides a conceptual model for the process of conducting design science research as well as a mental model for presenting the outputs of design science research.

The process model as defined by Peffers *et al*. (2006) consists of six activities, namely *problem identification and motivation*; *defining the objectives for a solution; design and development of an artifact*; *demonstration*; *evaluation*; and *communication*. The activities,

with the possible starting points of a design science research project are depicted in Figure 2.3.

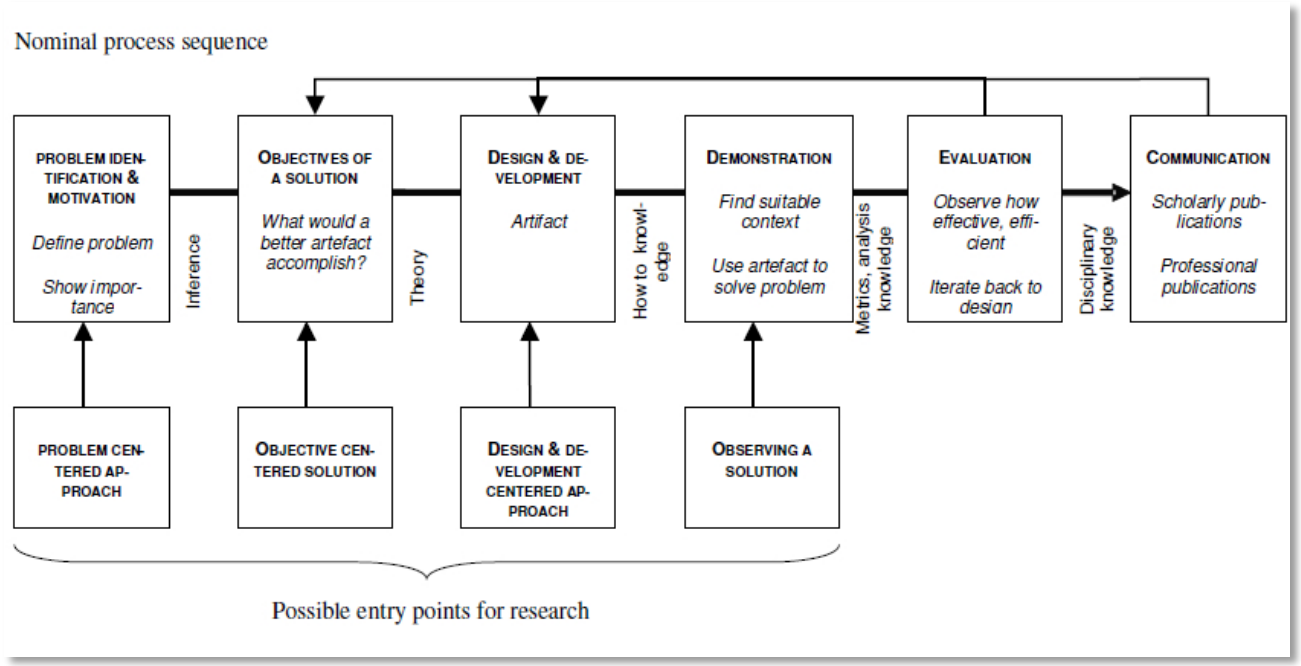


Figure 2.3: Design science research process (Peffers *et al*., 2006)

According to Peffers *et al*. (2006), there are four different possible starting points for conducting design science research, each representing a different research approach. Accordingly, a *problem-centred* approach starts at the identification of a problem and the justification for finding a solution to the problem. An *objective-centred* approach starts at the definition of the objectives that a solution should achieve. A *design and development-centred* approach starts at the design and development of an artifact that originated in an already existent artifact. A starting point for design science research might also originate from *observing a practical solution* that worked. The research follows a *problem-centred approach* and therefore the identification of the problem and the justification for finding a solution to the problem are considered as the starting point for the research. Figure 2.4 on the following page depicts an adaption of the DSRM of Peffers *et al*. (2006) as followed for this research study.

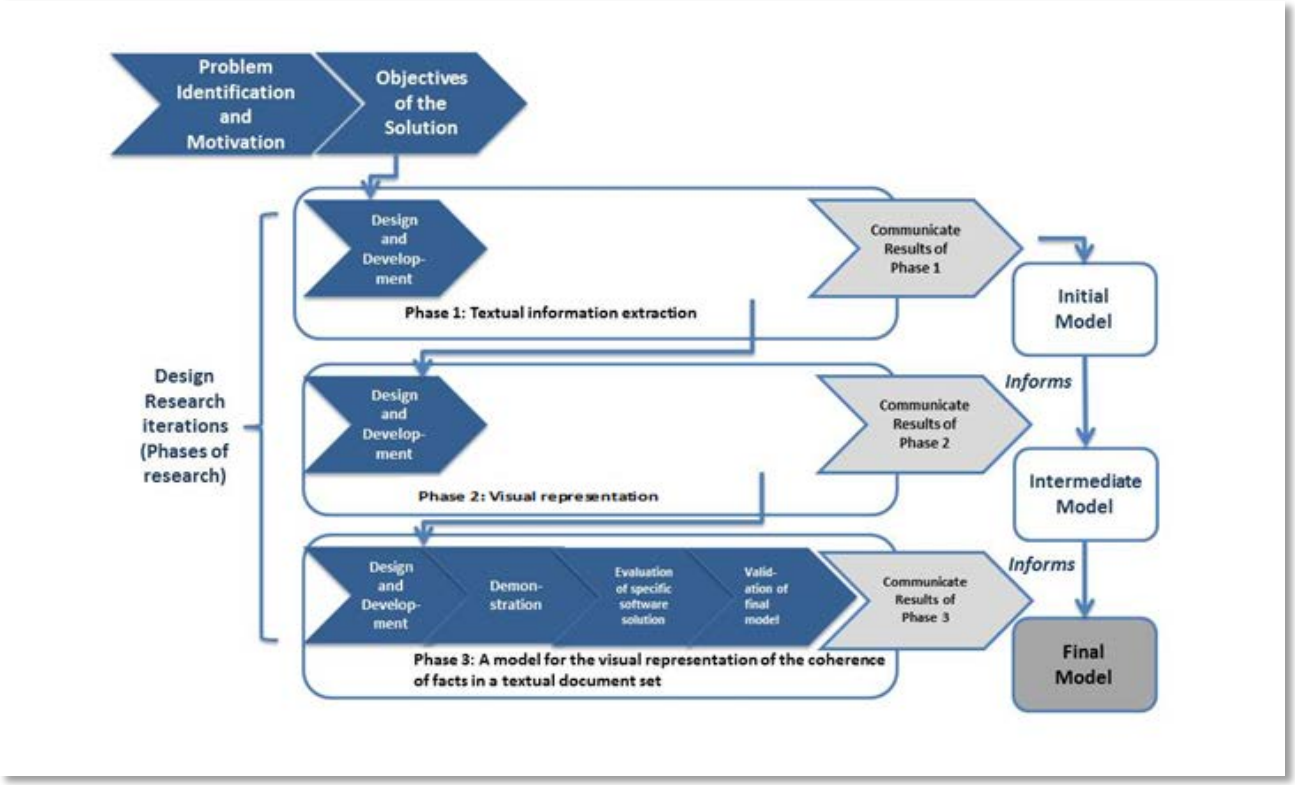


Figure 2.4: Adapted design science research process from Peffers *et al*. (2006)

In conducting the research study, three phases of designing, building and instantiating design science research artifacts were executed. These phases produced an *initial* textual information extraction model, *an intermediate* visual representation model and a *final* model for the visual representation of the coherence of facts in a textual document set. The identification and motivation of the problem as well as the defining of the objectives of the solution preceded these three phases.

The process of conducting the research is summarised as follows:

1. *Identify the problem and motivate the importance of the solution*

   During this activity, the problem is identified and the motivation for conducting the research is articulated. Identifying the problem allows for the development of an effective solution. By providing the motivation for the research, the researcher, as well as the research audience, is motivated to accept the research results and it also assists in understanding the researcher's understanding of the problem (Peffers *et al*., 2006).

   The problem statement of the research study is identified in Section 1.2.1 to Section 1.2.3 and a motivation for conducting the research is provided in Section 1.4 of the research study.

2. *Define the objectives of the solution*

   Peffers *et al*. (2006) postulate that the objectives that the solution should achieve, should be deduced from the problem definition of the research. Furthermore, Peffers *et al*. (2007) indicate that the objectives can be qualitative or quantitative. For the purpose of the research study, only qualitative objectives are set out. These objectives are listed in Section 1.2.5. The corresponding research phases addressing each objective are depicted in Figure 1.2.

3. *Phase 1 of the model development: Textual information extraction*

   During Phase 1 of the development of the model for visually representing the coherence of facts in a textual document set, an *initial information extraction* model is designed and developed.

   a. *Design and develop the initial model*

      An *initial* model for the extraction of information from a textual document set is designed, developed and instantiated during this activity. The model is deduced from an instantiation of a specific solution to extract and store information from a textual document set. The design and development of the *initial* information extraction model is discussed in detail in Chapter 3.

   b. *Communicate results of Phase 1*

   The results of this phase are communicated in Section 3.3 of this dissertation as well as in a conference article (Engelbrecht *et al*., 2016).

4. *Phase 2 of the model development: Visual representation*

   During Phase 2 of the development of the model for visually representing the coherence of facts in a textual document set, an *intermediate* visual representation model is designed and developed.

   a. *Design and develop the intermediate model*

      An *intermediate* model for the visual representation of the coherence of facts in a document set is designed, developed and instantiated during this activity. The model is deduced from an instantiation of a specific software solution to visually represent the coherence of facts. The design and development of the *intermediate* visual representation model is discussed in detail in Chapter 4.

   b. *Communicate results of Phase 2*

   The results of this phase are communicated in Section 4.3 of this dissertation as well as in a journal article (Engelbrecht *et al*., 2015) and conference articles (Engelbrecht *et al*., 2014, 2016; Engelbrecht & Botha, 2015).

5. *Phase 3 of the model development: Final information visualisation*

   During Phase 3 of the development of the model for visually representing the coherence of facts in a textual document set, a *final* visual representation model is designed and developed.

   a. *Design and develop the specific final software solution*

      A *final* model for visually representing the coherence of facts in a textual document set is designed and developed by deducing the model from an instantiation of the specific *final* software solution; the *final* solution being the conglomerated instantiation of the *initial* and *intermediate* solutions. Consequently, the *final* model is a conglomerated model consisting of the *initial* and *intermediate* models from Phase 1 and Phase 2.

   b. *Demonstration of the specific final solution*

      According to Peffers *et al*. (2006), this activity in the design science research process involves the demonstration of the efficacy of the produced artifact to solve the problem, which includes a number of activities such as its use in experimentation, simulation, a case study, a proof or any other appropriate activity. Peffers *et al*. (2006), in addition, state that the demonstration of the artifact requires an effective knowledge of how to use the artifact to solve the problem.

      For the purpose of the research study, the *final* solution instantiation was demonstrated to experts in the field of Anglo-Boer War history as well as to IT experts. The demonstration mainly consisted of a walkthrough of scenarios indicating how the coherence of facts in the Huyser collection of Anglo-Boer War testimonies were visualised (the visual representation), but also included a demonstration of how the information was extracted and used to populate the relational database feeding the user interface (the information extraction process).

      A simple feedback questionnaire (described in Section 5.1) was presented to the audience at the demonstration in the form of five-point Likert scale statements. The statement set was designed specifically to be used as input to the evaluation activity to determine the following:

      - whether design principles and guidelines were applied to the design of the visual representation;

      - whether the objective to identify components that can be combined into a model to extract facts and transform unstructured textual data into a structured form was achieved;

- whether the objective to identify components that can be combined into a model to visually represent the coherence of facts contained in a structured data form was achieved.

The demonstration of the specific *final* software solution is discussed in detail in Section 5.1.1.

c. Evaluation of the specific final software solution

During the evaluation activity of the design science research process, the researcher observes and measures whether the artifact provide a solution to the identified problem (Peffers *et al*., 2006). According to Peffers *et al*. (2006), objectives of the solution could be compared with the observed results from the artifact demonstration during this activity. Furthermore, according to them, the evaluation could be a comparison of the functionality provided by the artifact to the set objectives or objective quantitative performance measures.

A number of design evaluation methods as listed in Table 2.4 can be employed. The specific methods applied to the research study are marked in the table.

Table 2.4: Applying Hevner et al.'s (2004) design evaluation methods

Design evaluation methods		**Applied to the research study**
1. Observational	Case study, field study	
2. Analytical	Static analysis, architecture analysis, optimisation, dynamic analysis	**X**
3. Experimental	Controlled experiment, simulation	
4. Testing	Functional testing, structural testing	
5. Descriptive	Informed argument, scenarios	**X**

An *analytical analysis* was conducted on feedback questionnaires received from participants after demonstrations of the *final* software solution and model during interviews, and a *descriptive* analysis was conducted by demonstrating scenarios of visually representing the coherence of facts in a document set (the Huyser collection of Anglo-Boer War testimonies) to the participants. The demonstrated scenarios indicated how the coherence of facts in the Huyser collection were visualised (the interface), but also included a demonstration of how the information was extracted from the textual set and used to populate the relational database feeding the user interface.

At the end of the evaluation activity, the researcher can decide to iterate back to the design and development activity to improve the solution artifact, or to leave the artifact improvements to subsequent projects. For the purpose of the

study, due to time constraints, improvements to the artifact are considered as future work and therefore no reiteration was done.

For the purpose of the research study, the evaluation of the solution instantiation entails the following:

- comparing the objectives of the research with the functionality provided by the solution instantiation. This is done by carefully including statements in the feedback questionnaire in order to evaluate whether the research objectives of the study was met. The result is a decision (based on a pivot table analytical analysis of the feedback) on whether the resulting artifact provides a useful tool for gaining insight into domain facts contained in a set of textual records and a decision on whether the resulting artifact visually represents the coherence of facts contained in a textual record set;

- reflecting on the feedback provided by the participants of the demonstration interviews conducted in the previous activity. The result is a list of possible improvements.

The evaluation of the resulting artifact is discussed in detail in Section 5.1.2.

d. Validation of the final model

After the demonstration of the specific *final* software solution instantiation, the deduced general *final* model was presented to a participant audience comprised of the same IT experts as in the *final* solution demonstration. As in the case of the *final* solution demonstrations, the participants were required to complete a simple feedback questionnaire in the form of five-point Likert scale statements (described in Section 5.2.1). The statements where designed in order to determine by means of a pivot table analysis of the feedback (described in Section 5.2.2) whether the *final* model consists of a conglomeration of the *initial* information extraction and *intermediate* visual representation models; whether the *final* model represents components for visually representing the coherence of facts in a textual document set; and whether the model was informed by the instantiations of the *initial* and *intermediate* solutions.

e. Communicate the research results

Peffers *et al*. (2006) advise that the problem, the importance of the problem, the artifact, the utility and novelty of the artifact, the rigour of its design and its effectiveness should be communicated to an audience.

In addition to the research dissertation, the following was communicated to an academic audience through conference proceedings and in an academic journal publication:

- The principles and guidelines to guide the design of an information visualisation view provided in Section 4.1.3 (Engelbrecht *et al*., 2014, 2015).

- A proposed feedback questionnaire that could be presented to the artifact demonstration participants as presented in Section 5.1.1 (Engelbrecht *et al*., 2015). The feedback questionnaire presented to participants during demonstrations of the *final* solution instantiation, originated from this proposed feedback questionnaire.

- Design science research as an appropriate approach to Information Visualisation research (Engelbrecht *et al*., 2014, 2015).

- The design of the visual representation of the coherence of facts contained in a record set articulated in Section 4.2.1 (Engelbrecht & Botha, 2015). The instantiation of visual representation solution was based on this proposed design.

- The *final* model for visually representing the coherence of facts in a textual document set, and therefore the final results of the study (Engelbrecht *et al*., 2016).

The preceding academic publications were subject to academic scrutiny through peer review processes. In addition to the peer-reviewed academic publications, the content of the research study was also presented at the SAICSIT 2014 Master's and Doctoral Symposium (SAICSIT 2014 Master's and Doctoral Symposium, 2014) and at the FASTAR/Espresso Workshop 2014 (Engelbrecht, 2014).

Having discussed the research paradigm, methodology and approach followed by the study, the reliability and validity of the study is considered in the next section.

2.4 RELIABILITY AND VALIDITY

In order to ensure reliability and validity of the research, the study subscribed to the seven guidelines for conducting design science research as proposed by Hevner *et al*. (2004). These guidelines as well as its application to the study were discussed in Section 2.2.1.4. In addition, to ensure that a rigorous and reproducible research process was followed, an adaption of the referenced design research process model provided by Peffers *et al*. (2006) was followed. The adapted process was discussed in Section 2.3.

The results of the study as well as the research methodology and method followed were communicated to an academic audience at 13th Community Informatics Research Network (CIRN) conference by means of a peer reviewed conference article (Engelbrecht *et al*.,

2016). Reliability and validity of the research were therefore added to the research by means of a peer review process.

Having outlined the overall research approach to achieving the solution, the next chapter provides the design, development and instantiation of the *initial* model to extract information from a textual document set.

CHAPTER 3: DEVELOPMENT OF THE INITIAL MODEL – TEXTUAL INFORMATION EXTRACTION

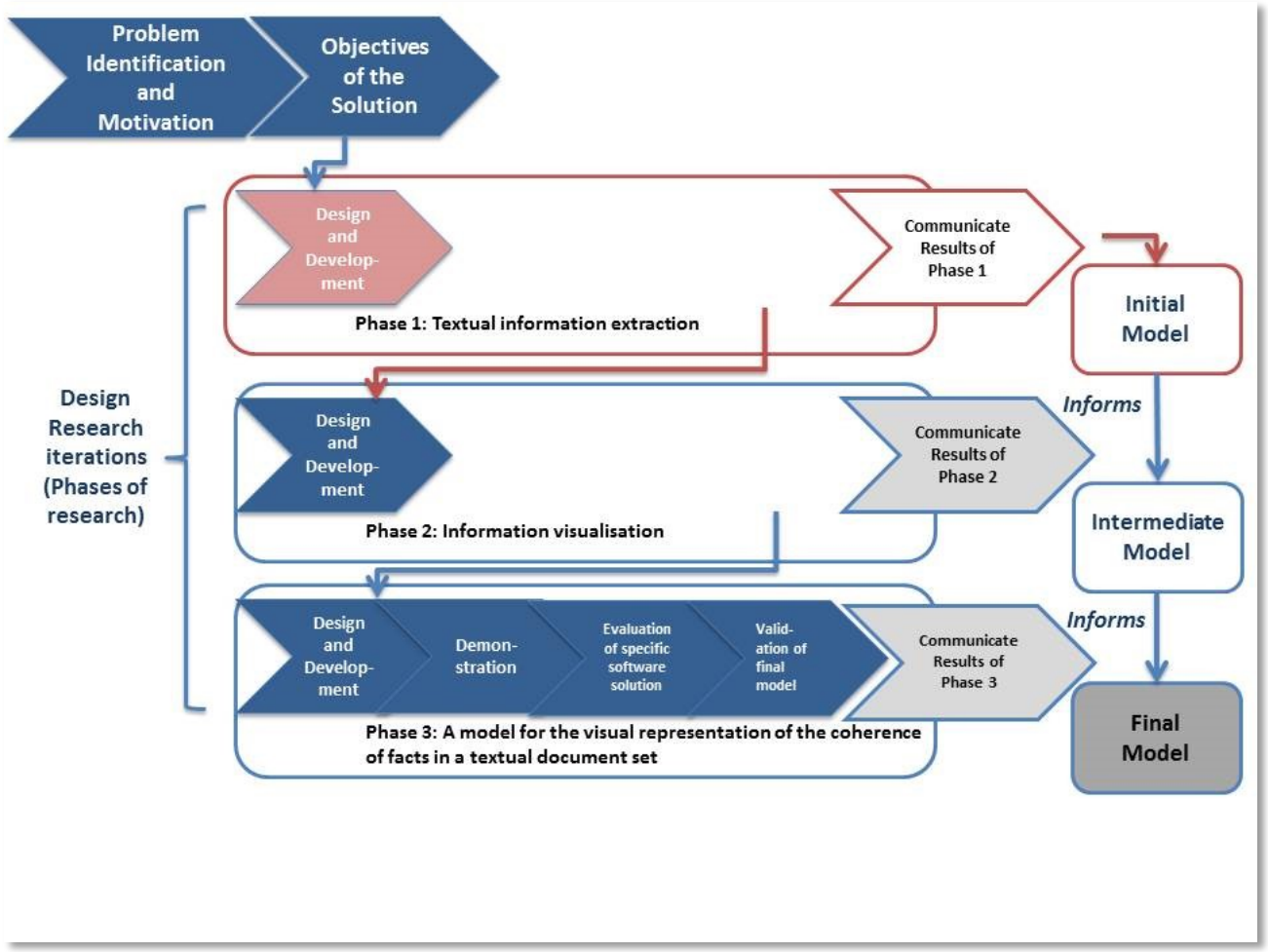


Figure 3.1: Current phase in research – textual information extraction

As depicted in Figure 3.1, Chapter 3 describes the development of the *initial* model for extracting information from a set of textual documents. The development of the model is informed by the design and instantiation of a specific *initial* information extraction solution. The design and development process starts with a discussion of literature relevant to the extraction of information from textual sources in order to ground the solution in literature (Section 3.1). The development of the *initial* model is subsequently informed by instantiating a proposed *initial* solution (Section 3.2). The chapter is concluded with a summary (Section 3.3)

3.1 LITERATURE RELEVANT TO THE EXTRACTION OF INFORMATION FROM TEXTUAL SOURCES

In order to ground the *initial* information extraction solution and model in literature, a review of literature regarding information extraction from textual sources is discussed in this section. The literature review focus on acquiring theoretical knowledge on what information extraction is (Section 3.1), which pre-processing tasks take place (Section 3.1.1), the available tools and techniques to extract information from text (Section 3.1.2) and the technologies available to store the extracted information (Section 3.1.3).

The review involved an extensive search on *Google Scholar* for related academic articles using keywords such as *information extraction*, *text mining*, *machine learning*, *text extraction, information extraction tools* and *information extraction techniques*.

Various definitions for *Information Extraction* (IE) exist in the literature, including the following:

- The process of retrieving information from textual sources in order to identify entities such as people, places and time and to identify the relationships between them (Gupta & Lehal, 2009).
- The task of populating a relational database schema with values detected in textual documents (Fagin, Reiss, Kimelfeld, & Vansummeren, 2016).
- The discovering of useful information from unstructured of semi-structured text contained in textual documents (Rai & Murari, 2015)
- The task of automatically extracting structured information from unstructured of semi-structured natural language in the form of text (Dang & Ahmad, 2015).
- The process of transforming unstructured information embedded in texts into structured data to enable further processing on the data (Jurafsky & Martin, 2014).
- The identification and extraction of specific relations of information from text (Schmitz, Bart, Soderland, & Etzioni, 2012).

For the purpose of the study, *Information Extracting* is considered to be the extraction (either manually or automatically) of information from unstructured textual records and the population of a database schema with the extracted information. Furthermore, it is derived from the various definitions of IE that unstructured textual records is considered to be text containing natural language like English where no processing has been done to extract information into structured records. These records describe the information in the textual records, enabling further processing of the information. Vijayarani *et al*. (2015) elaborate on the difference between structured and unstructured data postulating that structured data resides in fixed fields in a record or file, whereas unstructured data is the opposite of structured data and does not reside in a traditional relational database.

From the preceding discussions of *Information Extraction* it is concluded that a preliminary process for extraction information from textual records can be laid out as depicted in Figure 3.2.

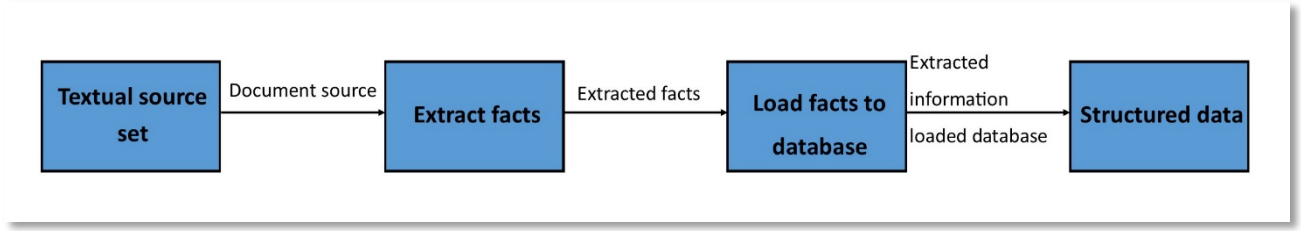


Figure 3.2: Preliminary information extraction process

The remainder of this section is laid out as follows. In Section 3.1.1 tasks taking place before information is extracted from textual document is discussed. Methods and tools identified from literature for the extraction of the information from the textual documents is discussed in Section 3.1.2. The storage of the extracted information is considered in Section 3.1.3. The section on literature relevant to the extraction of information from textual sources is concluded in Section 3.1.4

3.1.1 Preprocessing tasks on textual sources prior to information extraction

Various authors (Gupta & Lehal, 2009; He & Grishman, 2015; Vijayarani, Ilamathi, & Nithya, 2015; Zhang, Chen, & Liu, 2015) propose that preprocessing tasks usually take place on the textual sources before the information is extracted.

The following are some of the common preprocessing tasks identified by various authors (Grimmer & Stewart, 2012; Ravi & Ravi, 2015; Uysal & Gunal, 2014; Vijayarani *et al*., 2015):

- *Named Entity Extraction (NER).* Entities describing identifiable names such as persons and organisations are automatically detected and extracted from text (Cybulska & Vossen, 2011; Fagin *et al*., 2016; Jurafsky & Martin, 2014; Ravi & Ravi, 2015; Strötgen, Gertz, & Popov, 2010; Yang, Torget, & Mihalcea, 2011). Jurafsky and Martin (2014) provide a generic list of possible entity types that can be populated by an NER preprocessing task including people, organisations, locations, geo-political entities like countries and states, facilities and vehicles. Examples of software tools for performing names entity extraction on textual records are listed in Table 3.1.

- *Stemming* (Grimmer & Stewart, 2012; Ravi & Ravi, 2015; Uysal & Gunal, 2014; Vijayarani *et al*., 2015). The goal of stemming is, according to Vijayarani *et al*. (2015), to reduce the words in a textual source to their base forms in order to save time and computational memory space. As an example they use the words connect, connected, connection, connections and connecting to indicate that all these words

can be stemmed to the word *connect*. Vijayarani *et al*. (2015) in addition state that stemming algorithms can be classified into three groups of algorithms including *truncating methods, statistical methods and mixed methods.* Examples of software tools for reducing word in textual records to their base forms available are listed in Table 3.1.

- *Removal of stop words* (Grimmer & Stewart, 2012; Ravi & Ravi, 2015; Uysal & Gunal, 2014; Vijayarani *et al*., 2015)*. According to Grimmer and Stewart (2012)* it is words that appear more than 99% in the text as well as uncommon words appearing less than 1% in the text. Vijayarani *et al*. (2015) state that *stop words* should be removed from the textual source since these words make the text collection larger than it is and these words don't contribute to the meaning contained in the text; automatic information extraction therefore don't need these words. Uysal and Gunal (2014) state that stop word removal improves the efficiency of the subsequent information extraction tasks. Examples of software tools for removing stop words from textual records are listed in Table 3.1.

- *Correction* of spelling mistakes (Ravi & Ravi, 2015; Yang *et al*., 2011) in order to improve the quality of the text. An example of a software tool for spell checking documents is including in Table 3.1.

- *Clustering/Categorising.* The goal of the task is the classify textual documents into related categories (Dang & Ahmad, 2015; Grimmer & Stewart, 2012; Kaushik & Naithani, 2016). Dang and Ahmad (2015) as well as Kaushik and Naithani (2016) state that the classification is based on the content of the individual textual documents. The advantage of clustering, according to Dang and Ahmad, resides in that thousands of documents in an organisation can be organised in an effective manner. It is therefore a useful technique of organising large amounts of documents. An example of a software tool for clustering/categisring textual documents is included in Table 3.1

- *Tokenisation* (Ravi & Ravi, 2015; Uysal & Gunal, 2014). The task involves the tokenisation of a sentence into words, phrases, symbols and other meaningful tokens by removing punctuation marks (Ravi & Ravi, 2015). The result is a preliminary list of useful information. An example of a software tool available for the tokenisation of sentences, especially tweets is included in Table 3.1.

Preprocessing tasks including named entity extraction, stemming, removal of stop words and uncommon words, correction of spelling mistakes and textual source clustering/categorising are of particular interest when unsupervised or semi-supervised information automatic extraction methods (the different information extraction methods are discussed in Section 3.1.2.1) are employed. The tasks improve and enhance the subsequent automatic information extraction tasks. The reasons for using a supervised, manual method for the purpose of the research study to extract the facts from the textual source, is presented in Section 3.1.2.1. It is therefore suggested that for the purpose of the

study, with the exception of correcting spelling mistakes, the listed preprocessing tasks will not be considered in the *initial* information extraction solution. Preprocessing tasks relevant to the preparation of the textual documents before the information is manually extracted is identified during the instantiation of the initial solution. These tasks are discussed in Section 3.2.1.

Taking into account from the preceding discussion that preprocessing tasks usually takes place before information is extracted from the textual documents (for both manual and automatic extraction methods), it is proposed that a *document preparation* component is added to the preliminary information extraction process depicted in Figure 3.2. As a result, the proposed main components of the information extraction process is depicted in Figure 3.3. These components are used as input to the development of the *initial* information extraction model described in Section 3.2. The process is supported by Kaushik and Naithani (2016) who postulate that extracting information from text consists of a *preprocessing step*, an *information extraction* step and an *analysis* step. For the purpose of the study the visual representation of the coherence of the facts in the data is regarded as the analysis step.

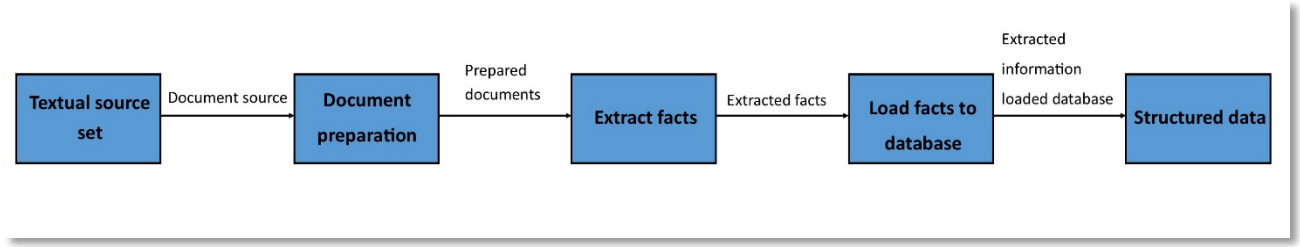


Figure 3.3: Proposed main components of the information extraction process

Having proposed main components of the information extraction process, the next section considers methods and tools identified in the literature to extract facts from the prepared textual document set.

3.1.2 Information extraction methods and tools

Having proposed main components of the information extraction process, this section considers possible methods (Section 3.1.2.1) and tools (Section 3.1.2.2) to extract information from textual documents. At the end of each section reasons for choosing a particular method or tool is provided.

3.1.2.1 Information extraction methods

Mooranian *et al*. (2011) categorise the information extraction methods into supervised, semi-supervised and unsupervised methods and consider semi-supervised and unsupervised methods to be automatic extraction methods. Supervised methods are considered by them as manual methods.

Grimmer and Stewart (2012), on the other hand, divide automatic information extraction methods only into two categories, namely supervised and unsupervised methods. Grimmer and Stewart's (2012) supervised method category corresponds to Mooranian *et al*.'s (2011) semi-supervised category. For the purpose of the research study to also include manual extraction methods, the information extraction categorisation of Mooranian *et al*. (2011), namely supervised, semi-supervised and unsupervised methods, is considered as follows:

- *Supervised extraction methods* are considered by various authors (Cybulska & Vossen, 2011; Grimmer & Stewart, 2012; Gupta & Lehal, 2009; McCallum, 2005; Yeung & Jatowt, 2011) as the traditional, manual method of extracting information from textual sources. Tremblay, Parra and Castellanos (2015) regard the manual method of reading through the text and assigning parts of texts to defined code as the most common technique to extract information from text.

  However, manual methods of extracting information from textual sources are difficult and time consuming, especially when large set of textual documents are used (Cybulska & Vossen, 2011; Grimmer & Stewart, 2012; Yeung & Jatowt, 2011). According to McCallum (2005), manual information-extracting methods are also prone to errors as a result of the human effort in the extraction process.

  Various software tools for *manually* extracting information from textual records are listed in Table 3.1.

- *Semi-supervised extraction methods*, also referred to as *supervised training* methods by Grimmer and Stewart (2012) and Jurafksy and Martin (2014), involve the hand coding (manually identifying themes and text related to the themes) of a sample from the textual source set, subsequently using the hand-coded sample to train or supervise statistical models (Grimmer & Stewart, 2012; Gupta & Lehal, 2009). The hand-coded extracted information is then used an input to the statistical extraction model, resulting in a set of similar information from the whole set. The quality of the extracted information therefor depends on the initial training set. This dependency is, as in the case of manual extraction methods, prone to errors due to human involvement in the process.

  According to Grimmer and Stewart (2012) and Gupta and Lehal (2009), the purpose of training a statistical model is to automatically extract information (based on the training) contained in the remainder of the textual document. An adequate amount of documents should however be available as a training set.

  The steps for *semi-supervised* information extraction methods are defined by Grimmer and Stewart (2012). Firstly, a training set is manually constructed. The training set is then applied to statistical methods. The last steps involve validating and assessing the performance of the training method.

When only a limited number of documents are available as training data for the automatic extraction methods, and approach known as *distant supervision* can be followed (Jurafsky & Martin, 2014; Roller, Agirre, Soroa, & Stevenson, 2015; Roller & Stevenson, 2015). Instead of relying on the source document set for producing training data, an external knowledge source is used. The knowledge source contains related entity pairs, which are used to extract sentences from the textual source which contain the same pairs. These sentences are then used as input into the automatic extraction method. As an example Roller and Stevenson use a drug and side effect database as an external sources and then use the drug-side effect pairs in the database to extract sentences in drug reports with these pairs. The extracted sentences are then used as training data for automatic information extraction from the set of drug reports. According to Roller and Stevenson (2015) the distant supervision approach however do not perform as well as manually trained data.

Various software tools for using *semi-supervised* information extraction methods are listed in Table 3.1.

- *Unsupervised extraction methods* involve finding information in the textual source set without imposing theme categories, with the additional benefit of identifying unknown areas of interest (Grimmer & Stewart, 2012). Named Entity Extraction (NER) (Fagin *et al*., 2016; Jurafsky & Martin, 2014; Ravi & Ravi, 2015), *clustering/categorising* (Dang & Ahmad, 2015; Grimmer & Stewart, 2012; Kaushik & Naithani, 2016) *, tokenisation* (Ravi & Ravi, 2015; Uysal & Gunal, 2014)*, topic modeling* (Grant *et al.*, 2015) and *word cloud representations* (Grant *et al*., 2015) are considered to be some of the available unsupervised extraction methods*. Topic models* represents a set of topics contained in a large set of documents, therefor presenting a summary of the content in the set (Grant *et al*., 2015). *Word clouds* are used to visualise the frequency of data contained in textual sources (Devendorf, 2012; Gan *et al*., 2014; Grant *et al*., 2015; Lu & Liu, 2011). An example of a word cloud from Grant *et al*. (2015) is depicted in Figure 3.4 on the following page. *Topic models* and *word clouds* are discussed in in Section 4.1.4.1. Various software tools for using *unsupervised* information extraction methods are listed in Table 3.1.

Figure 3.4: Word cloud example from Grant *et al*. (2015)

Semi-supervised and unsupervised extraction methods are particular useful when a large set of documents is available for information extraction (Cybulska & Vossen, 2011; Grimmer & Stewart, 2012; Gupta & Lehal, 2009; Yeung & Jatowt, 2011). Only for smaller sets of documents should the manual extraction method be considered.

Related to semi-supervised and unsupervised extraction methods, are the concepts of *text-mining* and *data-mining*. According to McCallum (2005) and Vijayarani *et al*. (2015), both *text-mining* and *data-mining* techniques use automatic (machine-learning) methods to extract data from text, but the difference is in the origin of the input (input is either from a database or from textual information). Kaushik and Naithani (2016) in particular refers to the information being extracted either from structured data (*data-mining*) or from unstructured data (*text-mining*). During the process of *data-mining*, the textual source is selected and inserted into a database from where the information mining takes place. In the case of *text-mining*, the mining is not performed on a database, but on the originating textual source. The result of the *text-mining* is afterwards stored in a database (McCallum, 2005; Vijayarani *et al*., 2015).

For the purpose of the study it was decided to use a manual, supervised method to extract the facts from the textual document set. In the process of selecting the approach to follow, the decision was based on the following considerations:

- Semi-supervised and unsupervised information extraction approaches are most appropriate when handling large sets of documents (Jurafsky & Martin, 2014; Tremblay *et al.*, 2015; Verma, Ranjan, & Mishra, 2015). The research study however provide a solution for a finite set of not more than 50 documents as included in the list of delineation and assumptions of the study in Section 1.5.

- Using semi-supervised and unsupervised information extraction approaches does not provide fully accurate results (Kaushik & Naithani, 2016) and are less accurate than manually extracting information from text (Recupero *et al*., 2015). For the purpose of the study it is considered as important that all important and especially relevant facts containing meaning should be extracted.

- Semi-structured and unsupervised approaches to extract information from text is not a trivial task (Rai & Murari, 2015). In addition, He and Grishman (2015), suggest that current automatic information extraction tools do not provide adequate guidance for the user building a new extraction system. The study attempts not to find the best solution, but a suitable solution. Therefore it is suggested that, grounded on pragmatism (Creswell, 2014; Gray, 2013), the method which requires the least effort could be considered – as long as it works. Tremblay *et al*. (2015) state that the supervised, manual method of reading though the textual source and assigning parts of the texts to predefined code is the most common and simplest method of extracting information from text.

- Semi-structured and unsupervised approaches extract entities such as persons, locations, organisations and events without a proper understanding of the text ant putting these entities into context with the rest of the text; requiring further processing of the information (Dang & Ahmad, 2015). It is suggested by the study that in order to extract only facts relevant and important to the Anglo-Boer War, a proper understanding of the context of the facts is of importance.

- Although automatic information extraction by using *distant supervision* could be an option, the requirement of an external database containing entity pairs related to the Anglo-Boer War could not be found. Distant supervision was therefore not considered as an extraction method.

- During the presentation of the proposed research study at the SAICSIT 2014 Master's and Doctoral Symposium (http://www.saicsit2014.org) and at the FASTAR/Espresso Workshop 2014 (http://www.fastar.org), the researcher was advised that the automatic extraction of the suggested Anglo-Boer War facts would require extensive research. Furthermore, the researcher was advised that combining semi-automatic or automatic information extraction research with the intended visualisation research would require an amount of research that is beyond the scope of a master's degree study.

  Due to the scope of the study and the amount of research required to combine semi-supervised or unsupervised information extraction methods with the intended visualisation research, the supervised manual information extraction method as identified in the literature study is proposed as extraction method for the purpose of the research study to extract the facts from the textual sources.

- The study follows a pragmatic approach as discussed in Chapter 2. According the philosophical view of pragmatism, truth is based on what works at the time to address the research problem (Creswell, 2014; Gray, 2013). For the purpose of the study, the intend is not necessarily to come up with the best possible information extraction solution, but rather to come up with an approach that achieves the goal of extraction facts from the Huyser collection of Anglo-Boer War testimonies. Future research can focus on improving the information extraction process.

In this section of the literature review, methods to extract information from textual data were discussed. The next section discusses tools identified in the literature that can be used to extract information from textual sources.

3.1.2.2 Information extraction tools

Information extraction tools identified in the literature are discussed in this section. In addition, a short description of some of the output formats for storing the extracted information is provided. Table 3.1 lists some of the information extraction tools and some of their uses, as identified in the literature. A comparative analysis of information extraction tools are not part of this research study, but rather investigating and identifying a suitable tool which can be used to extract facts from a textual document set.

Therefore, Table 3.1 does not present an extensive list of information extraction tools and all of the functionalities provided by them. The list was constructed by updating a list of tools found during the review of literature on information extraction methods discussed in Section 3.1.2.1. The list provide only examples of information extraction functions as well as some of the tools that can provide these functionalities. Future research can be aimed at providing a more comprehensive list of tools with all the functionalities provided by each.

Table 3.1: Example list of information extraction functions and tools

Information extraction function	Tools	References
Stemming	Intelligent Miner	(Dang & Ahmad, 2015)
	Text Analyst	(Dang & Ahmad, 2015)
	Snowball	(Ravi & Ravi, 2015)
Stop words	Intelligent Miner	(Dang & Ahmad, 2015)
	Text Analyst	(Dang & Ahmad, 2015)
Correction of spelling mistakes	GNU Aspell	(Ravi & Ravi, 2015)
Clustering/Categorisation	Intelligent Miner	(Dang & Ahmad, 2015; Kaushik & Naithani, 2016)
Tokenisation	TweetMotif	(Ravi & Ravi, 2015)
Tagging temporal expressions	TARSQI	(Ling & Weld, 2007; Strötgen *et al*., 2010; Yeung & Jatowt, 2011)
	Evito	(Ling & Weld, 2007)

Information extraction function	Tools	References
Named-entity recognition (NER)	ANNIE	(Strötgen *et al*., 2010)
	GATE	(Rusu, Fortuna, Mladenic, Grobelnik, & Sipo, 2009)
	Inxight	(Strötgen *et al*., 2010)
	Stanford Named Entity Recogniser	(Yang *et al*., 2011)
	OpenCalais	(Rusu *et al*., 2009)
Automatic topic modelling	Stanford Topic Modeling Toolbox	(Gretarsson *et al*., 2012)
	MALLET	(McCallum, 2005; Yang *et al*., 2011)
	GATE	(McCallum, 2005; Mooranian *et al*., 2011)
	MinorThird	(McCallum, 2005)
	Road-runner	(McCallum, 2005)
Unsupervised and semi-supervised information extraction	R tool	(Castillo & Gesmann, 2011; Keim *et al*., 2012; O'Madadhain, Fisher, Smyth, White, & Boey, 2005)
	OpenNLP	(Schreiber, Kraft, & Zündorf, 2016)
	DiscoTEX	(Rai & Murari, 2015)
Manual extraction	ATLAS.ti	(Franzosi, Doyle, McClelland, Putnam Rankin, & Vicari, 2013; Friese, 2011; Smit, 2014; Wong, Liu, & Bennamoun, 2012)
	MAXQDA	(Franzosi *et al*., 2013)
	NVivo	(Franzosi *et al*., 2013)

As discussed in Section 3.1.2.1, the manual information extraction method is chosen as suitable for the study. From the list of information extraction examples listed in Table 3.1, it is consequently considered that ATLAS.ti, MAXDA or NVivo might be suitable for the study to extraction information from the textual source. For the purpose of the study ATLAS.ti is considered as a suitable software tool to manually extract the required facts from the textual sources for the following reasons:

- It was suggested to the researcher by colleagues as an easy to use and appropriate tool for manually extracting information from textual sources.
- An initial investigation into the use of ATLAS.ti showed that facts can be extracted from documents while keeping track of the position where the fact occurred in the document. This is useful for visually representing the fact as well as enabling the user to refer to the context of the fact in the original document. According to the visualisation framework of Eppler and Burkhard (2007) it is important that the context of the information should be considered in the visualisation.

- An initial investigation into the use of ATLAS.ti showed that the extracted facts can be stored in an Extensible Markup Language (XML) file format. This is considered as useful for the purpose of the study since XML format that can easily be used to store the data in a relational database (Cybulska & Vossen, 2011; Don *et al*., 2007; Ling & Weld, 2007; Strötgen *et al*., 2010; Yeung & Jatowt, 2011).

- As discussed in Section 3.1.2.1 when choosing a suitable information extraction method the study did not considered the best solution, but a suitable solution. The decision to use ATLAS.ti is therefore based on to the pragmatic philosophy of truth which indicates that truth is based on what works at the time to solve the research problem (Creswell, 2014; Gray, 2013). Future research can be focused on finding the best software tool for the purpose of the study.

The information extracted from textual sources needs to be stored in some format. In the next section data storage technologies is discussed.

3.1.3 Storage of the extracted information is considered

Having discussed information extraction tools and software in the previous sections, this section considers technologies for storing the extracted data. Data in its structured form is mainly stored in Extensible Markup Language (XML) files (Emerson, Churcher, & Deaker, 2013; O'Madadhain *et al*., 2005) or relational database systems (Ku, Nguyen, & Leroy (2012); O'Madadhain *et al*., 2005).

The comparative study of Agarwal and Srivastava (2016) however provides a more extensive list of different storage systems, comparing each of them. In addition the advantages and disadvantages of each is also provided. The storage systems discussed by them are:

- *Flat file system*. The data is stored in an ordinary file on a computer, where the data in individual records are separated by delimiters such as commas. Although this system is simple and easy to understand, it can be difficult and taking long to access the required data. (Agarwal & Srivastava, 2016)

- *Relational database system*. The data is stored in a digital system which is based on the relational model of data. The data is accordingly to the model stored in tables consisting of rows and columns. The system uses the Standard Query Language (SQL) to easily access and modify the data. Performance of this system however reduces when large data sets and a large number of tables are used. (Agarwal & Srivastava, 2016)

- *Distributed database system*. The data is also stored in a digital system as in the case of the relational database system. The data is however distributed data over

computers on a network into parts according to groups of information. Data is better protected, more reliable and load performance of large data sets are improved. A distributed database system is however complex, difficult to maintain and requires skilled database administrators. (Agarwal & Srivastava, 2016)

- *Parallel database system*. The data is stored in the same manner as in a distributed database system. The difference however is that the groups of data is not according to information areas, but according by performance considerations. The input and output of these systems utilise multiple Central Processing Unit (CPU) s and memory devices to increase speed and reliability of especially large data sets. The start-up is however high and it is a complex system requiring a high number of resources.(Agarwal & Srivastava, 2016)

- *Object oriented database system*. Data is also stored digitally, not as ordinary data such as integers and characters, but as objects. Each object describes data in terms of its attributes (such and integers and characters) as well as methods that define the behaviour of the data. Some of the advantages of this system includes improved reliability, flexibility and reusability of programming code. Some of the disadvantages include query optimisation inefficiency, lack of support to view the data, lack of support for security and querying the data is difficult and costly. (Agarwal & Srivastava, 2016)

- *Object relational database system*. Data is stored as in a similar manner as in a relational database system. The data is however not stored according to the relational model as rows and columns in a table, but according to an object orientated database model. The goal of this system is to overcome the gap between the relational database system and the object orientated system. Increased costs related to the complexity of this system is one of the disadvantages. (Agarwal & Srivastava, 2016)

- *Extensible Markup Language (XML) based database* system. Data is stored in a specific XML format which can be easily queried (using the XQuery programming language), transformed and exported to various systems. Since it is a standard format and many organisations have a lot of data in the XML format, it can be easily exported to other current database systems. The data is human readable and therefore less expertise is required to access the data than in the case of a relational database system. Some of the advantages include slower query speed, data quality can be a problem since data integrity is not guaranteed and the entire database have to be loaded before any data can be viewed. (Agarwal & Srivastava, 2016)

- *Data warehouse database system.* Data from various data sources are stored in a central repository. Since multidimensional data is stored in one place, it allows for the data to be viewed in a generalised and consolidated view. These systems allows for enhanced business intelligence systems supporting organisational decision making. Warehouse systems requires high maintenance and are therefore costly. Complexity of these systems are also very high.

A comparison between the preceding list of database storage systems provided by Agarwal and Srivastava (2016) is provided in Table 3.2.

Table 3.2: Comparison between database storage systems (Agarwal & Srivastava, 2016)

	Flat file	Relational	Distributed	Parallel	Object oriented	Object relational	XML based	Warehouse
I/O Cost	High	Moderate	High	High	Moderate	Moderate	Very Low	Very High
Setup Cost	Low	Moderate	High	High	High	High	Very Low	Very High
Modularity	Very Low	Low	High	High	High	High	Low	Very High
Concurrency	Very Low	Moderate	High	High	High	High	High	Very High
Design Complexity	Very Low	Low	High	High	Low	Low	High	Very High
Security	Very Low	Low	High	High	Low	Moderate	Low	Very High
Flexibility	Very Low	Low	Moderate	High	High	High	Low	Very High
Maintenance	Low	Low	High	High	Low	Moderate	High	Very High
Reliability	Low	Moderate	High	High	High	High	Low	Very High
Reuse and Sharing	Very Low	Low	Moderate	High	High	High	Low	Very High
Intelligence	None	Very Low	Moderate	Moderate	Low	Low	Moderate	Very High
Throughput	Very Low	Low	High	High	Moderate	High	Low	Very High
Speed	Very Low	Low	Moderate	High	Moderate	Moderate	Low	Very High

For the purpose of the study, a relational database storage system is chosen to store the extracted information in.

The decision is based on the following considerations:

- A relational database system makes it possible to easily query and manipulate data using Standard Query Language (SQL). This makes subsequent processing of the extracted information a simple task.
- The amount of extracted information is relatively small, therefore not requiring more complex database systems that can specifically handle large volumes of data, as listed in Figure 3.2.

- Although the setup cost is not as low as in the case of flat file and XML based systems, the user's own work experience makes setting up a relational database system a trivial task.

- The development of the *intermediate* visual representation model requires that a visual representation solution must be built programmatically. The researchers own skillset as a developer lead to the decision to build the solution on the architecture depicted in Figure 4.19, which includes the use of a relational database system.

- A manual information extraction method was selected in Section 3.1.2.1 as suitable for the *initial* information extraction solution. In Section 3.1.2.2, ATLAS.ti was consequently selected as a suitable software tool performing the extraction. ATLAS.ti provides the functionality to store the information in the XML based storage system. The XML based system makes transport and export of data possible by means of a standard format. Exporting the XML based database to a relational database is therefore, based on the researcher's skillset, a trivial task.

Having provided a literature review of information extraction relevant literature in order to define possible main information extraction components and to guide the selection of a suitable information extraction method and tool, a summary of this section is provided next.

3.1.4 Summary

Concluding remarks on literature relevant to the extraction of information from textual sources are given in this section. The discussion started with an investigation into various views in literature of *Information Extraction* and a statement that for the purpose of the study, *Information Extraction* is considered to be the extraction (either manually or automatically) of information from unstructured textual records and the population of a database schema with the extracted information. From the consideration of what *Information Extraction* is, it was suggested that the *textual source set*, *extraction of facts*, *loading of facts into a database* and the resulting *structured data* are possible components (depicted in Figure 3.2) of an *Information Extraction* process.

Possible preprocessing tasks taking place on the textual sources before the information is extracted were discussed next. The discussion was motivated by various sources in literature suggesting that preprocessing tasks usually take place prior to the extraction of the information. As a result the *preparation of documents* was added to the possible information extraction components, resulting in the main component model depicted in Figure 3.3.

After main components of an information extraction process was identified from literature, various extraction methods and tools was considered next. As a result it was decided with motivations that a manual information extraction process could be followed using the ATLAS.ti software tool.

Using the possible identified main components, methods and tools as input, the next section discuss the instantiation of the specific information extraction solution. During the process of instantiating the solution an *initial* information extraction model is developed.

3.2 A MODEL FOR THE EXTRACTION OF INFORMATION FROM A TEXTUAL DOCUMENT SET

The purpose of the research study requires that a model for the visual representation of the coherence of facts in a textual document set be developed. In this section an *initial* model to initially extract the facts from the textual document set is developed by instantiating the possible main components identified in Figure 3.3. During the instantiation of the solution the components of the model are refined and sub-components identified.

According to Hevner *et al*.'s (2004) design science research guidelines, design science is a search process whereby the researcher searches for an effective solution to a problem. In searching for a solution (grounded on the literature review in Section 3.1) to extract information from a textual document set and populating a database schema with the information, the research study identified that the following steps were required in the process:

1. Prepare the documents for information extraction (document preparation).

2. Extract the information contained in the documents (extract facts).

3. Populate the database schema with the extracted information (load facts to database).

4. Identify possible errors in the information extraction process, and iterate back to Step 2 until the errors are considered to be insignificant (discovered during the solution instantiation process).

The steps were identified during the literature review (Section 3.1 and Figure 3.3) and refined during the instantiation of the solution. The information extraction steps are discussed in the next sections, starting with the preparation of the textual documents.

3.2.1 PREPARE TEXTUAL DOCUMENTS FOR INFORMATION EXTRACTION

During the preparation of the document set for information extraction, the following processes were identified:

1. *Acquire* the document set.

   The starting of the information extraction process requires an appropriate document set to be acquired. The document set can either be in the form of written documents or any other form of data that can be transformed into a document set such as audio and video files containing indigenous knowledge in an oral format. However, the

document set must exhibit the characteristics listed in Section 1.2.1 to be considered as a textual source.

For the purpose of the study, as the primary data set was not viable, the Huyser collection of Anglo-Boer War testimonies was used because it exhibited the prescribed characteristics to be used as an alternative source,

Since the Huyser collection of Anglo-Boer War testimonies is stored at the South African National Archives, the pages of the testimonies had to be individually photographed. Figure 3.5 presents an example of such a photographed page.

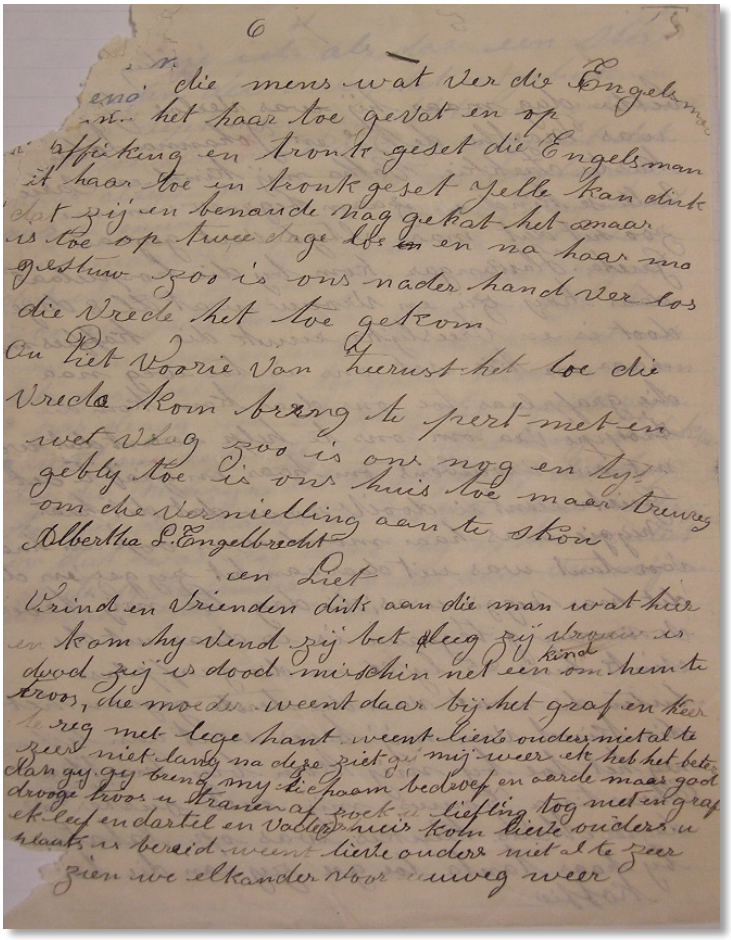


Figure 3.5: Example of a photographed page

2. *Transcribe* the text contained in the document set into a digital text format.

   Having acquired the document set, it needs to be transcribed into digital text files to allow for further processing. Transcribing each document from the original format (e.g. digital images or audio or video files) is a time-consuming task which

furthermore requires the transcriber to decipher the written or spoken words and sentences into an understandable format.

3. *Translate* the language of the text into English for the use of a larger audience.

   In order to reach a larger audience, the language of the text contained in the Huyser collection was translated from Afrikaans into English. The following steps were followed in the translation process:

   a. Employ *Google translator* to create an intermediate English translation.

   b. *Inspect and correct* the intermediate English translation. It was found during the process that using *Google translator* to translate the documents resulted in a number of incorrect translations such as the following:

      i. Incorrect translations of the double negative used in Afrikaans. Refer to Oosthuizen (1998) for the use of the double negative in Afrikaans.

      ii. Literal translations of names and surnames. The surname Vermaak for instance was translated to its literal meaning, which is Entertainment.

      iii. Incorrect interpretations where a word can have different meanings. The word *ryp* in Afrikaans for instance can mean ripe or frost. In some cases, the author used the word in the context of frost, but *Google translator* translated the word as ripe.

      iv. In some cases some translated sentences did not make sense at all, for instance: “play to be buried” should be “set aside to be buried”.

   c. *Remove* defamatory words from the translation and *replace* the names of certain individuals with codes in order to protect the identities of individuals. The need for this step is required by the ethical issues to the research as discussed in Section 1.6.

Having described tasks for preparing documents in a document set for information extraction, the next section articulates a process for extraction information contained in a textual document set

3.2.2 EXTRACT THE INFORMATION CONTAINED IN A TEXTUAL DOCUMENT SET

A proposed process for extracting information from a textual source set is provided in this section. For the purpose of the research study it was decided, based on literature in Section 3.1.2.1, to use a supervised, manual information extraction method on the textual document set (the Huyser collection of Anglo-Boer War testimonies).

During the literature review on information extraction tools (see Section 3.1.2.2), ATLAS.ti was identified with reasons as a possible manual information extraction tool. As part of the research study, a trial version of the tool was downloaded and tried out. As a result, the tool was identified as a user-friendly, easy-to-use information extraction tool. Furthermore, it was found that the facts extracted from the document set could be easily exported to an XML data representation format. The extracted information exported to an XML file can then be used to upload the data into the database system. Based on the trial evaluation of the software, it was decided to employ ATLAS.ti as information extraction tool.

In order to be able use ATLAS.ti for extracting the required information from the textual document set, it is necessary to map some of the concepts used in the research study with ATLAS.ti concepts. A summary of the mapping of these concepts is provided in Table 3.3.

Table 3.3: Mapping of study concepts to ATLAS.ti concepts

Research study concepts	ATLAS.ti concept	Description
Document	Primary document	The source set consists of a number of textual documents. Each document contains the testimony of a person.
Fact	Code	Factual information (Chen & Floridi, 2013) described in a document set.
Fact occurrence	Quote	Each occurrence of a fact in the document set.

In order to extract information from the textual document set, the following procedure was followed using the ATLAS.ti software (the procedure is highlighted here to give the reader an insight into what the manual extraction method involves):

1. Import a textual document (termed as a primary document in ATLAS.ti) containing one of the testimonies into ATLAS.ti by using the primary document loader functionality.

2. Select the imported document for information extraction.

3. Highlight the text in the document that describes a fact (termed as a code in ATLAS.ti) as depicted in Figure 3.6 on the following page. The highlighted fact occurrence is termed as a quote in ATLAS.ti.

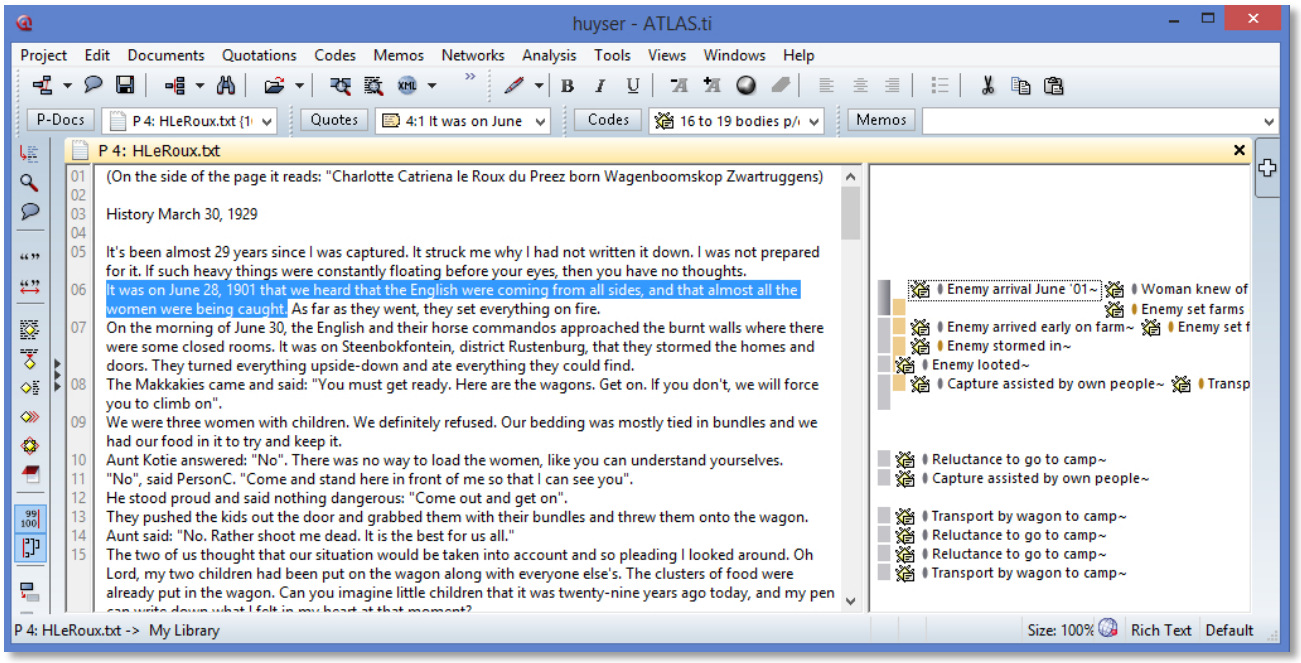


Figure 3.6: Selecting factual information using ATLAS.ti

4. By right-clicking on the highlighted text, the user can either create a new fact (code) for the highlighted text (quote), or can indicate that the highlighted text represents an occurrence of an already identified fact.

5. After importing and extracting the facts from all the documents in the document set, the user can export the extracted information into an XML data representation format by using ATLAS.ti's *Export HU to XML* functionality. The export XML file contains the following information:

    - the filenames of all the textual documents;
    - all the facts contained in the textual document set;
    - each occurrence of a fact in the document set, indicating its start position and end position in the text.

A snippet of the XML data representation created during the information extraction of the Huyser collection of Anglo-Boer War testimonies is depicted in Figure 3.7 on the following page.

```
 <q name="Some prayed and cried and othe.." id="q9_80" au="Super" cDate="2015-07-04T12:54:14" mDate="2015-07-04T12:54:14" loc="138 @ 11, 178 @ 11!"/>
 <q name="She then called for her uncle .." id="q9_81" au="Super" cDate="2015-07-04T12:54:44" mDate="2015-07-04T12:54:44" loc="1 @ 36, 135 @ 37!"/>
 <q name="We had to struggle to get some.." id="q9_82" au="Super" cDate="2015-07-04T12:56:47" mDate="2015-07-04T12:56:48" loc="55 @ 61, 176 @ 61!"/>
</quotations>
</primDoc>
<primDoc name="CEVanStaden.txt" id="pd_10" loc="doc_14" au="Super" cDate="2015-05-11T15:40:18" mDate="2015-05-31T19:11:52" qIndex="51" >
<quotations size="51" >
 <q name="We spent three days in the can.." id="q10_1" au="Super" cDate="2015-05-11T17:49:43" mDate="2015-05-11T17:49:43" loc="1 @ 1, 191 @ 2!"/>
 <q name="We fled because some people sa.." id="q10_2" au="Super" cDate="2015-05-11T17:59:55" mDate="2015-05-11T17:59:55" loc="1 @ 2, 96 @ 2!"/>
 <q name="I was eight days there when ru.." id="q10_3" au="Super" cDate="2015-05-11T18:00:40" mDate="2015-05-11T18:00:40" loc="1 @ 5, 208 @ 5!"/>
 <q name="Early the next morning (at 2 o.." id="q10_4" au="Super" cDate="2015-05-11T18:04:45" mDate="2015-05-11T18:04:45" loc="1 @ 6, 208 @ 6!"/>
 <q name="They fought for a few hours. T.." id="q10_5" au="Super" cDate="2015-05-11T18:07:28" mDate="2015-05-11T18:07:28" loc="1 @ 10, 185 @ 10!"/>
 <q name="The English then stormed into .." id="q10_6" au="Super" cDate="2015-05-11T18:07:59" mDate="2015-05-11T18:07:59" loc="1 @ 11, 51 @ 11!"/>
 <q name="finished everything. The wheat.." id="q10_7" au="Super" cDate="2015-05-11T18:08:16" mDate="2015-05-11T18:08:16" loc="57 @ 11, 253 @ 11!"/>
 <q name="One lot killed the chickens an.." id="q10_8" au="Super" cDate="2015-05-11T18:08:41" mDate="2015-05-11T18:08:41" loc="255 @ 11, 291 @ 11!"/>
 <q name="It was about at 8 o&amp;apos;clock in t.." id="q10_9" au="Super" cDate="2015-05-11T18:09:23" mDate="2015-05-11T18:09:23" loc="1 @ 12, 52 @ 14!"/>
 <q name="Just as the wagons were moving.." id="q10_10" au="Super" cDate="2015-05-11T18:10:07" mDate="2015-05-11T18:10:07" loc="1 @ 15, 41 @ 16!"/>
 <q name="The Boers, that day had shot m.." id="q10_11" au="Super" cDate="2015-05-11T18:10:36" mDate="2015-05-11T18:10:36" loc="1 @ 17, 174 @ 17!"/>
 <q name="killed pigs and chickens" id="q10_12" au="Super" cDate="2015-05-11T18:11:09" mDate="2015-05-11T18:11:28" loc="48 @ 18, 71 @ 18!"/>
 <q name="burned down houses" id="q10_13" au="Super" cDate="2015-05-11T18:12:00" mDate="2015-05-11T18:12:00" loc="77 @ 18, 94 @ 18!"/>
 <q name="We stayed for three days at re.." id="q10_14" au="Super" cDate="2015-05-11T18:15:18" mDate="2015-05-11T18:17:20" loc="1 @ 19, 106 @ 19!"/>
 <q name="They offloaded each family in .." id="q10_15" au="Super" cDate="2015-05-11T18:16:18" mDate="2015-05-11T18:16:19" loc="1 @ 20, 46 @ 20!"/>
 <q name="hunger" id="q10_16" au="Super" cDate="2015-05-11T18:17:42" mDate="2015-05-11T18:17:42" loc="80 @ 20, 85 @ 20!"/>
 <q name="the food they gave us was meag.." id="q10_17" au="Super" cDate="2015-05-11T18:18:29" mDate="2015-05-14T10:41:47" loc="115 @ 20, 147 @ 20!"/>
```

Figure 3.7: Snippet of XML data representation created by ATLAS.ti

Table 3.4 contains the number of documents, total number of unique facts extracted from the Huyser collection and the number of times the facts occurred in the documents.

Table 3.4: Number of documents, unique facts and the number of times facts occurred

Entity	Count
Documents	37
Unique facts	653
Number of times the facts occurred in the documents	2 137

Having discussed a process to extract the information from the textual documents, the next section discusses how to populate the database schema presented in Section 4.2.1.2 with the extracted information.

3.2.3 POPULATE THE DATABASE SCHEMA

The process of extracting factual information from textual documents was discussed in the previous section. The end-product of the process was an XML data representation of the extracted information as depicted in Figure 3.7. A process for populating the database schema with the extracted information is discussed in this section. As a result of the process of developing the information extraction artifact, it was discovered that, for efficiency, the data population process could be executed in three consecutive steps as follows:

1. Load the data-representing information about each document into the database. A simple Spring 3 web service utility named *loadDocs,* was developed for this purpose.

A description of the corpus is coded in the utility program. The corpus description is displayed in the visual interface in order to provide background information to the user. A code snippet of the utility program is displayed in Figure 3.8.

```
Project project =(Project) jaxbUnmarshaller.unmarshal( new File("c:\\louis\\tmp\\hu_testimonies.xml") );


for(PrimDoc primDoc: project.getPrimDocs().getPrimDocList()){

    String idAsString = primDoc.returnId().replace("pd_", "");
    Long id = Long.parseLong(idAsString);


    Document document = documentService.findOne(id);
    System.out.println(document);
    if (document == null){
        Author author = new Author();
        author = authorService.modify(author);
        document = new Document();
        document.setId(id);
        document.setFileName(primDoc.returnName());

        document.setAuthor(author);
        System.out.println("hier");
        documentService.modify(document);

        Corpus corpus = corpusService.findOne();

        if (corpus == null){
            corpus = new Corpus();
```

Figure 3.8: Code snippet of *loadDocs* utility

2. Load the contents of each document into the database. The location and document identifier of each text file containing the document contents are manually entered into a PHP (Hypertext preprocessor) script file named *texttodb.php*. A code snippet of the PHP script file is depicted in Figure 3.9.

```
$data = fread($h, filesize('AMCSnyman.txt'));
$escaped = pg_escape_string($data);
fclose($h);
$sql = "update document set content = '" . $escaped . "', numberofcharacters = " . strlen($data) ."   where id = 5";
$conn = pg_connect('host=localhost port=5432 dbname=textvisdemo user=textvis password=textvis');
pg_query($conn,$sql);


$h = fopen('AMEFourie.txt', 'r');
$data = fread($h, filesize('AMEFourie.txt'));
$escaped = pg_escape_string($data);
fclose($h);
$sql = "update document set content = '" . $escaped . "', numberofcharacters = " . strlen($data) ."   where id = 6";
$conn = pg_connect('host=localhost port=5432 dbname=textvisdemo user=textvis password=textvis');
pg_query($conn,$sql);


$h = fopen('Annoniem1.txt', 'r');
$data = fread($h, filesize('Annoniem1.txt'));
$escaped = pg_escape_string($data);
fclose($h);
$sql = "update document set content = '" . $escaped . "', numberofcharacters = " . strlen($data) ."   where id = 7";
$conn = pg_connect('host=localhost port=5432 dbname=textvisdemo user=textvis password=textvis');
pg_query($conn,$sql);
```

Figure 3.9: Code snippet of texttodb.php script file

3. Load the data representing all the facts, as well as data of their occurrences in the document set, into the database. A simple Spring 3 web service utility, named *loadInformation,* was developed for this purpose. A code snippet of the utility program is displayed in Figure 3.10.

```
Map<String, Quotation> quoteMap = new HashMap<String, Quotation>();

for(PrimDoc primDoc: project.getPrimDocs().getPrimDocList()){

    String idAsString = primDoc.returnId().replace("pd_", "");
    documentId = Long.parseLong(idAsString);
    System.out.println(primDoc.returnId().replace("pd_", ""));

    document = documentService.findOne(documentId);

    String str = "";
    if (document.getContent() != null){
        str = new String(document.getContent(), "UTF-8");
    }

    int index = str.indexOf("\n");

    Map<Integer, Integer> lines = new HashMap<Integer, Integer>();
    lines.put(1,0);
    int numberofLines = 1;

    while (index >= 0) {
        numberofLines++;
        index = str.indexOf("\n", index + 1);
        lines.put(numberofLines, index);

    }
```

Figure 3.10: Code snippet of loadInformation utility

Having prepared the textual documents, extracted the information in the documents and populated the database schema, the last step is to identify possible information extraction errors by means of inspection.

3.2.4 IDENTIFY POSSIBLE ERRORS INTRODUCED DURING THE INFORMATION EXTRACTION PROCESS

The last step in the information extraction process involves the identification of possible errors introduced during the information extraction process. According to McCallum (2005), manual extraction methods are prone to errors. Therefore, it is considered necessary to inspect the populated database schema for possible errors. However, it is not a trivial task to identify possible information extraction errors. After the visualisation interface has been built, it can, in addition, be used to identify possible information extraction errors by inspecting the information visualised in the interface.

The user can, for instance, identify the following errors:

- missing information such as occurrences of facts not identified;

- information that does not make sense. An occurrence of a fact, for instance, “they received bad meat”, suggests that the occurrence of “they received meat” is also expected;

- duplicate facts. It might be possible that during the manual extraction different codes for the same fact were created, for instance, “they received bad meat” and “the meat quality was poor” might both occur in the list of code list.

After possible errors are identified, the process iterates back to the *extract facts* component as depicted in Figure 3.11 on the following page to resolve the errors until the significant errors are resolved.

Components of a process to extract facts from textual data was discussed in this section. The process of discovering these components was initially informed by possible main components identified though a literature review in Section 3.1. By instantiating a solution, components of an *initial* information extraction model was developed. The model is discussed in the next section.

3.3 SUMMARY

An *initial* model for the extraction of facts and the transformation of unstructured textual data into a structured form is presented next by reflecting on the information extraction components identified in the previous section.

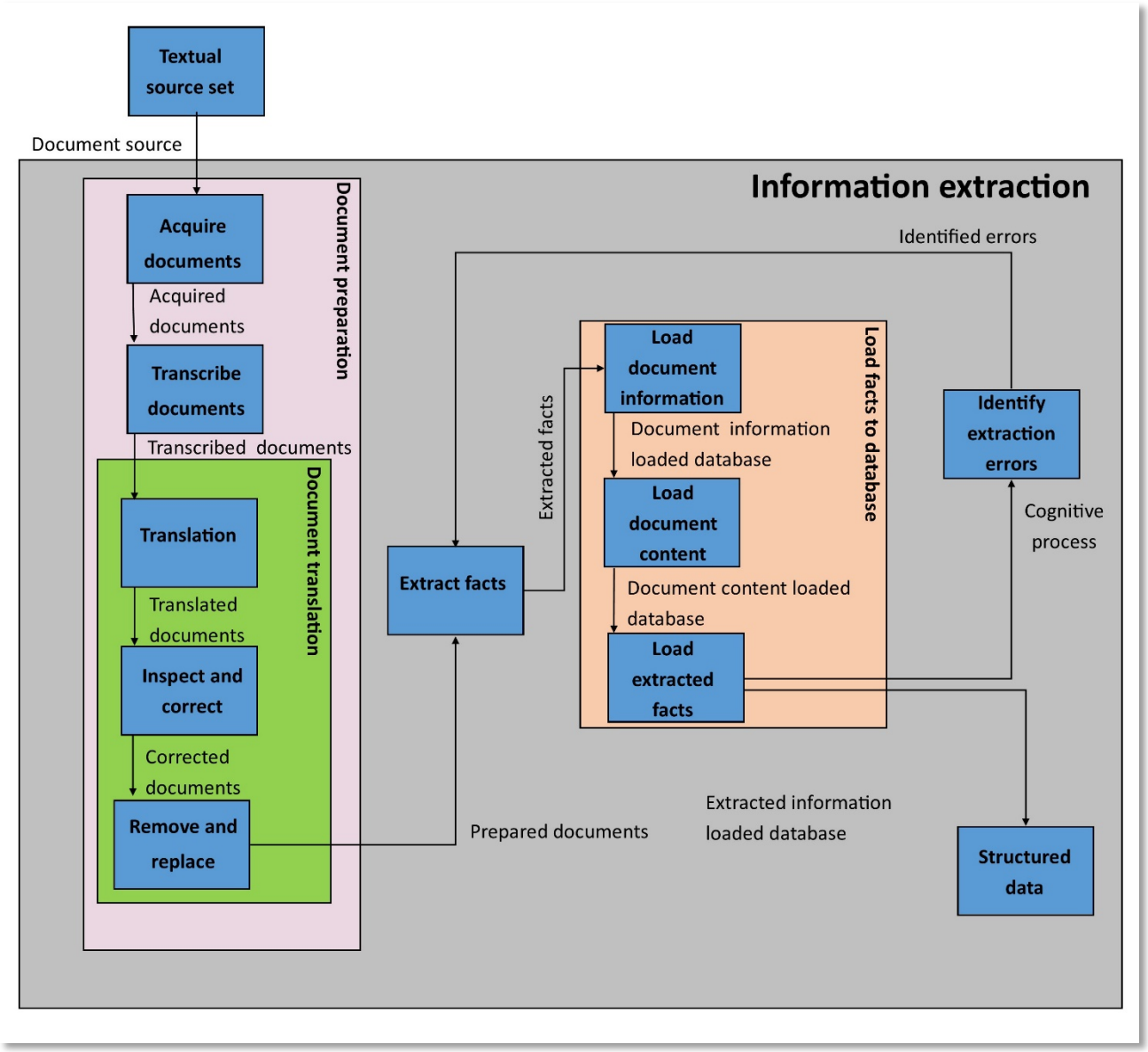


Figure 3.11: Initial model – extracting facts and transforming unstructured textual data into a structured form

Therefore, the following components can be combined into a model (depicted in Figure 3.11) to extract facts from a textual document set and to transform unstructured textual data into a structured form:

- *Textual source set*. This consists of the source of the data from which the documents are to be acquired. In the case of the Anglo-Boer War testimonies used in the study, the source is the Huyser collection. In the case of the indigenous knowledge, the source can be the NRS initiative.

- *Prepare documents* for information extraction.
    - *Acquire documents*. This consists of the collection of the documents from the source. The documents can be written documents or any other form of data that can be transformed into a document set, for instance, audio and video files containing indigenous knowledge in an oral format.
    - *Transcribe documents*. This involves the transcription of each document into a digital text file.
    - *Translate documents*. This involves the translation of each document into a language such as English in order to reach a larger reader group.
        - *Translate* using a translation tool such as *Google trans*lator.
        - *Inspect* for possible translation errors and *correct* the translation where required.
        - *Remove* defamatory words and *replace* names of individuals with code in order to protect their identities.
- *Extract facts* from the prepared textual source set. For the purpose of the research study, a manual method is described. Future research can be an investigation into employing unsupervised and semi-supervised information extraction methods.
- *Load* the extracted facts into a database for use by the visualisation interface of which the design is provided in Chapter 4.
    - *Load document information* describing each document into the database.
    - *Load document contents* of each textual document into the database.
    - *Load the extracted facts* from each document into the database.
- *Identify errors* during the information extraction process. It is proposed that the errors can be identified upon inspection of the extracted data set, or by using the visual representation to visually highlight possible errors. Errors are resolved by iterating to the *extract facts* component as depicted in Figure 3.11 to resolve the errors until the significant errors are resolved.

Having provided the design and development of an *initial* model to extract information from a textual document set in this section, the next chapter presents the design and development of an *intermediate* model to visually represent the coherence of facts in a document set.

CHAPTER 4: DEVELOPMENT OF THE INTERMEDIATE MODEL – INFORMATION VISUALISATION

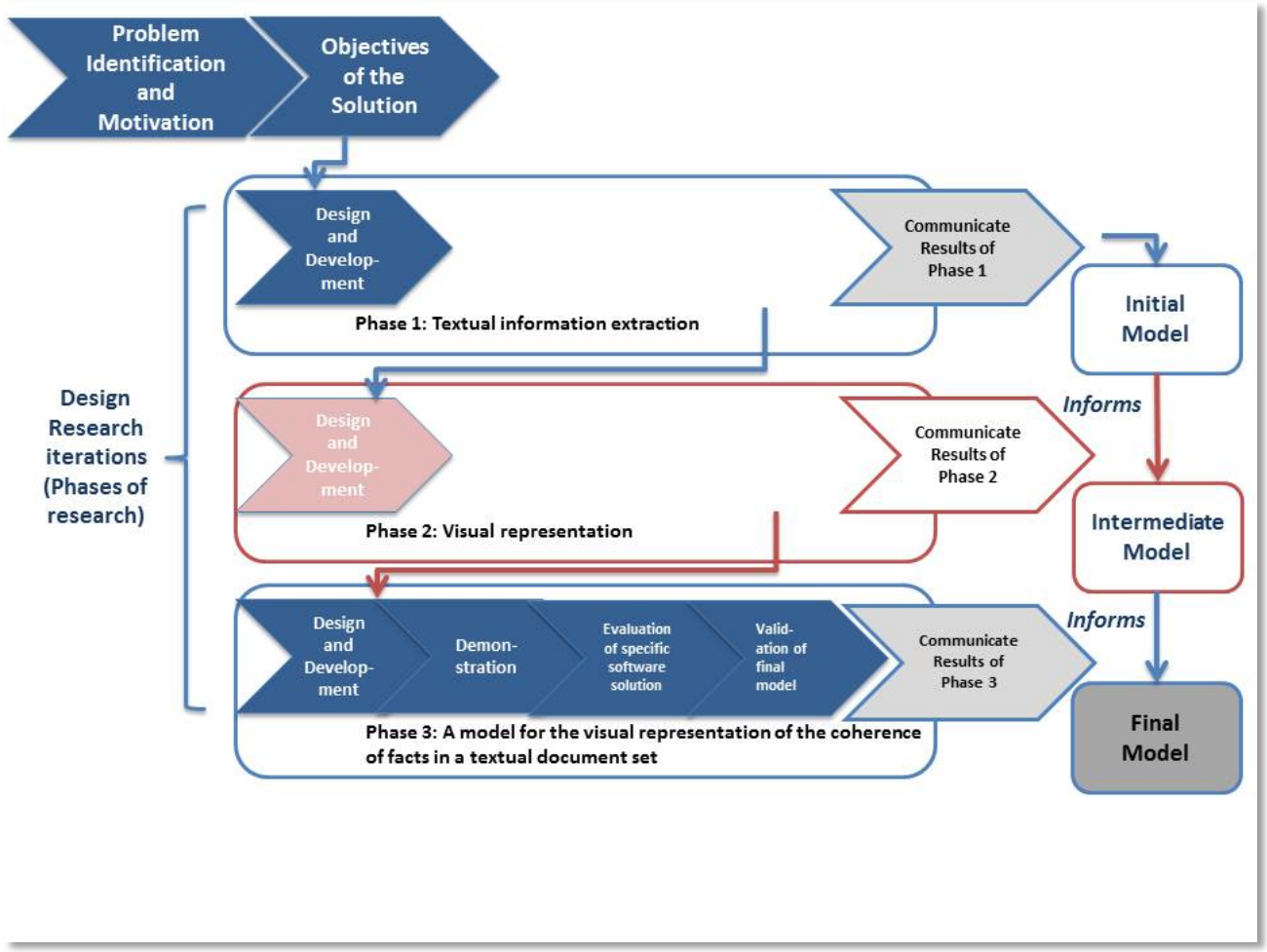


Figure 4.1: Current phase in research – visual representation

As depicted in Figure 4.1, Chapter 4 describes the design and development of the *intermediate* visual representation model. The development of the model is informed by the design and instantiation of a specific *intermediate* visual representation solution. In order to ground the solution in literature, the design and development process starts with a discussion of literature relevant to the visualisation of information coherence (Section 4.1). The development of the *intermediate* model is subsequently informed by instantiating a proposed *intermediate* solution (Section 4.2). The chapter is concluded with as summary (Section 4.3).

4.1 LITERATURE RELEVANT TO THE VISUALISATION OF INFORMATION COHERENCE

In order to ground the *intermediate* visual representation solution and model in literature, a review of literature relevant to the visualisation of the coherence of facts is discussed in this section. The literature review focus on similar work (Section 4.1.1) and acquiring theoretical knowledge on background information concerning visualisation (including Ackoff's hierarchy to explain the difference between information visualisation and knowledge visualisation) concepts (Section 4.1.2), design principles and guidelines to guide the visualisation design (Section 4.1.3) and information visualisation tools and techniques (Section 4.1.4).The review involved an extensive search on *Google Scholar* for related academic articles using keywords such as *information visualisation*, *coherence visualisation*, *visualisation*, *visualisation of text, visualisation design, visualisation principles, visualisation evaluation, visualisation tools* and *visualisation techniques*.

The remainder of this section is laid out as follows. Section 4.1.1 discuss work similar to the research study. Working definitions for the concepts of *Visualisation*, *Knowledge Visualisation*, *Information Visualisation* and *Document Visualisation* are derived from literature in Section 4.1.2. Literature on *Visualisation* design principles is discussed in Section 4.1.3 in order to construct a set of *Information Visualisation* design principles and guidelines. In Section 4.1.4, literature with regard to tools and techniques available for implementing the *Information Visualisation* interface design is discussed. The chapter on literature relevant to the visualisation of information coherence is concluded in Section 4.1.5.

Having outlined what the discussion of literature relevant to the visualisation of information coherence entails, the next section discuss work similar to this study.

4.1.1 Similar work

Work similar to this research study is discussed in this section. For the purpose of the research study, work is considered to be similar when the focus of the work is on the visualisation of textual documents to visually represent the connections between them. In this regard, the work of Gan *et al*. (2014), Ku *et al*. (2012), Gretarsson *et al*. (2012), Devendorf (2012) and Luo, Yang, Krstajic, Ribarsky and Keim (2012) is considered as similar and is shortly discussed:

- *Gan et al. (2014).* Their work is similar to the research study in the sense that they also visually represent information contained in a textual document set. The purpose of their work is to visually represent the connections between entities in a textual document set. This research study attempts to visualise the coherence of facts contained in the document set, therefore, requiring a different information extraction method combined with a different visualisation method. Examples of tools that can be used to visually represent the connections between entities contained in a set of textual documents are, in addition, provided by Gan *et al*. (2014).

- *Ku et al. (2012).* The textual analysis of similar crimes (TASC) system of Ku *et al*. (2012) is considered similar to this research study because it also aims to visually represent textual information contained in a document. The difference between the work of Ku *et al*. (2012) and this research study resides in the type of information extracted from a textual document and also what is being visualised. This research study is concerned with visually representing the coherence of facts contained in a textual document set, whereas Ku *et al*. (2012) extract specific entity types, such as weapons and people, from each document and then visually represent documents that contain the same entities, without taking into account what is being described by these entities. Ku *et al*. (2012), in addition, show that document visualisation techniques can also be applied to the field of crime investigations.

- *Gretarsson et al. (2012).* The similarity between textual documents in a set is highlighted by Gretarsson *et al*. (2012), by visually representing their similarities with regard to the topics extracted from individual documents. A circular node-and-link graph visualisation is implemented to visually represent the similarities. A visual representation of topic similarity between textual documents, however, only indicates that the documents describe similar themes. Whether the same facts are contained in the documents is not taken into account. This research study visually represents the coherence of facts in a document set that goes beyond visually indicating whether the documents describe the same themes or entities as in the case of Ku *et al*. (2012).

- *Devendorf (2012).* The work of Devendorf (2012) is similar to the work of Gretarsson *et al*. (2012) in the sense that the author also visually represents the similarity between topics contained in a textual document set. However, the visualisation is more complex and complete because it visually represents three relationships contained in textual documents, by dividing the circular graph into three concentric circles. The concentric circles visually represent three types of relationships, namely meta-data relationships, relationships between topics and items and relationships between topics and ternary data. As in the case of Gretarsson *et al*. (2012), the work of Devendorf (2012) is concerned with similarity between themes in a set of textual documents; this research study is concerned with visually representing the coherence of facts contained in the set of textual documents.

- *Luo et al. (2012).* The focus of Luo *et al*.'s (2012) work is on the extraction and visualisation of events that exist in a document set where each document has a timestamp. Luo *et al*. (2012) define an *event* as "an occurrence that happens at a specific time and draws continuous attention" (p. 3). Events contained in a set of textual documents are visualised as a river that progresses over time. Each event is shown as a bubble where the width of each bubble represents the period in which the event is being considered and the height represents the number of documents in which the event occurs. Luo *et al*.'s (2012) work is different from the work of Gretarsson *et al*. (2012), Devendorf (2012), Gan *et al*. (2014) and Ku *et al*. (2012) with respect to what is extracted and how it is being visualised. In addition, Luo *et al*.

(2012) state that extracting events from textual documents requires more human effort than for topic extraction. The work of Luo *et al*. (2012) is different from this research study in terms of what is being extracted (events vs. facts), and also *how* it is being visualised (occurrence of events vs. coherence of facts).

It is proposed from the preceding discussion that the research study is different from previous work in the literature with respect to *what* is extracted from textual documents and *how* the extracted information is visualised.

Section 4.1.3.4 on the Eppler and Burkhard (2007) *Knowledge Visualisation* framework states that the difference between *Knowledge Visualisation* and *Information Visualisation* lies in the answers to *what* is visualised, *why* it is visualised and *how* it is visualised. The research study considers the visualisation of document similarities as the objective (the why) of the Gan *et al*. (2014), Ku *et al*. (2012), Gretarsson *et al*. (2012), Devendorf (2012) and Luo *et al*.'s work. In addition to *what* information (facts) is visualised and *how* it is visualised (see Section 4.2.1.1 for design of the visualisation interface), the research study also differs from previous work in *why* (to make an assumption on the validity of the contained facts) the information is visualised.

Having discussed work similar to the research study, the next section provides working definitions of visualisation concepts.

4.1.2 Defining Visualisation, Knowledge Visualisation, Information Visualisation and Document Visualisation

This section provides literature on defining various working definitions of *Visualisation* concepts in order to provide subject background and to justify the development of an *Information Visualisation* software tool as a suitable insight-acquiring solution. As such, it is considered as important to provide a working definition for *Information Visualisation.*

There are currently various definitions of *Information Visualisation* in the literature, which include the following:

- The visualisation of information about the data itself (Chen *et al*., 2009).
- "All graphical means used to construct, assess, measure, convey or apply knowledge" (Eppler & Burkhard, 2007, p. 112).
- "Making data easier to understand using direct sensory experience" (Dix, 2013, p. 2).
- "A visual exploration tool that enables the user to interact with the visualized content and comprehend its meaning" (Chen, 2005, p. 12).
- "A key mechanism for gaining insights from text collections" (Mooranian *et al*., 2011, p.138).

- “The use of computer-supported interactive visual representations of abstract data to amplify cognition" (Card, Mackinlay, & Shneiderman, 1999).

- “Used for creating computer graphs or animations to present information, data, scientific results, or concepts to facilitate communications or decision making” *(Ku et al*., 2012, p. 467).

- “The study of transforming data, information, and knowledge into interactive visual representations” (Liu, Cui, Wu, & Liu, 2014, p. 1373).

- “The process of dissecting raw data, which by itself has little meaning and presenting it in a way that is no longer complex” (Figueiras, 2014, p. 2).

Considering that the concepts of *Data*, *Information*, *Knowledge* and *Wisdom* are ambiguous (Chen *et al*., 2009; Keller & Tergan, 2005; Liew, 2013; Nurnberger *et al*., 2009) and that there are various views on *Visualisation* (Chen, Floridi, & Borgo, 2013; Chen & Floridi, 2013; Keller & Tergan, 2005; Randjelovic & Popovic, 2011), consequently, the assumption is that the concepts of *Data Visualisation*, *Information Visualisation*, *Knowledge Visualisation* and *Wisdom Visualisation* are also ambiguous.
It is suggested that because *Information Visualisation* is proposed as an insight-acquiring solution for the study and since the concept of *Information Visualisation* is vague and ambiguous, a working definition of *Information Visualisation* that is relevant to the research study should be formulated. It is proposed that a working definition of *Information Visualisation* can be derived from the available literature by considering the following:

- a working definition of *Visualisation* derived from the literature (Section 4.1.2.1);

- working definitions of *Knowledge* and *Information* derived from the literature (Section 4.1.2.2);

- literature considering the difference between *Information Visualisation* and *Knowledge Visualisation* (Section 4.1.2.3).

A working definition of *Document Visualisation* is in addition discussed in Section 4.1.2.4 as it is considered as relevant to the study. In the next section, a working definition of *Visualisation* is derived by considering different views of *Visualisation* in the literature.

4.1.2.1 Working definition of *Visualisation*

The idea that Visualisation leads to an insight is central to various discussions of Visualisation (Dix, 2013; Eppler & Burkhard, 2007; Gretarsson *et al*., 2012; Keim *et al*., 2012). *Insight* is defined by Dove and Jones (2012, p. 2) as “something that is gained” and as “something that is experienced”.

On the other hand, Chen *et al*. (2013) state that *insight* is a vague concept and that the real purpose of Visualisation is to save time when completing user tasks involving data analysis

and knowledge dissemination. Some of the user tasks for which Visualisation can be employed to save time are making observations in data, facilitating external memorisation, simulating hypothesis and disseminating knowledge.

Despite the concept of *insight* being central to many Visualisation discussions, there are various definitions and views of what Visualisation is such as the following:

- " . . . cognitive tools aiming at supporting the cognitive system of the user" (Keller & Tergan, 2005, p. 2);
- ". . . computer-aided seeing information in data" (Chen & Floridi, 2013, p. 1);
- ". . . a set of technical means and methods that allow a clear graphical representation of data" (Randjelovic & Popovic, 2011, p. 1);
- ". . . a computer graphic effect to represent information and revealing relationships" (Kaushik & Naithani, 2016, p. 71)
- ". . . a study of transformation from data to visual representations in order to facilitate effective and efficient cognitive processes in tasks involving data" (Chen *et al*., 2013, p. 6).

The latter description by Chen *et al*. (2013) is considered the most complete description of Visualisation. The description is furthermore the most relevant to the research study. For the purpose of the research study, the definition provided by Chen *et al*. (2013) is adapted in order to present a working definition of Visualisation.

A working definition of **Visualisation** for this study is therefore as follows:

> *A field of study concerned with the transformation of data to visual representations, where the goal is the effective and efficient cognitive processing of data.*

Kaushik and Naithani (2016) describe the visualisation process (the transformation of data to visual representation part of the working definition) as consisting of:

1. data preparation;
2. data analysis and extraction and;
3. the mapping of the data to a target visualisation.

Step 1 and Step 2 is considered to be addressed by the *initial* information extraction model of the study (see Chapter 3). The mapping of the data to a target visualisation is considered as being address by this chapter – the development of an *intermediate* visual representation model.

Having derived a working definition of *Visualisation*, the next section derives working definitions of *Information* and *Knowledge* from available literature.

4.1.2.2 Working definitions of *Knowledge* and *Information*

Chen *et al*. (2009) postulate that *Visualisation* is concerned with the exploration of *Data* and *Information* and, in addition, consider *Data*, *Information* and *Knowledge* as the possible inputs and outputs of the visualisation process. Therefore, Chen *et al*. (2009) consider it important to have an understanding of the concepts of *Data*, *Information* and *Knowledge* and to be able to distinguish between them. Various authors suggest that it is not a trivial task to distinguish between them (Chen *et al*., 2009; Keller & Tergan, 2005; Liew, 2013; Nurnberger *et al*., 2009). Chen (2009) argues that although it is difficult to distinguish between these concepts, they are indeed distinctive. According to Liew (2013), defining these concepts and describing the relationships between them are separate issues.

In order to define these concepts, the Ackoff (1989) hierarchy of human understanding is considered. According to Kirsch, Hine and Maybury (2015), the hierarchy is used in a wide variety of disciplines to understand the process of acquiring knowledge. *Data*, *Information*, *Knowledge*, *Understanding* and *Wisdom* are presented as layers in the hierarchy of human understanding (depicted in Figure 4.2 on the following page). Ackoff (1989) argues that some form of processing is required in order to progress from one level to another. In order to obtain *Wisdom* the process starts off with meaningless *Data*. *Information* is acquired by extracting meaning from the *Data*. *Knowledge* is acquired next by some form of cognitive processing of the *Information*. An *Understanding* or intelligent processing of the *Knowledge* leads to *Wisdom*.

Nurnberger *et al*. (2009) consider the relationships between the concepts in the hierarchy to be “fuzzy” (p. 1), and propose that the “fuzziness” (p. 1) increases as one moves up the hierarchy. Accordingly, Nurnberger *et a*l. (2009) imply that it would be more difficult to distinguish between, for instance, *Knowledge* and *Information* than it would be to distinguish, for instance, between *Information* and *Data*. It is consequently suggested that it is also more difficult to distinguish between *Information Visualisation* and *Knowledge Visualisation* than to distinguish between *Data Visualisation* and *Information Visualisation*. For the purpose of the study, this section accordingly provides working definitions for *Information Visualisation* and *Knowledge Visualisation* to distinguish between them.

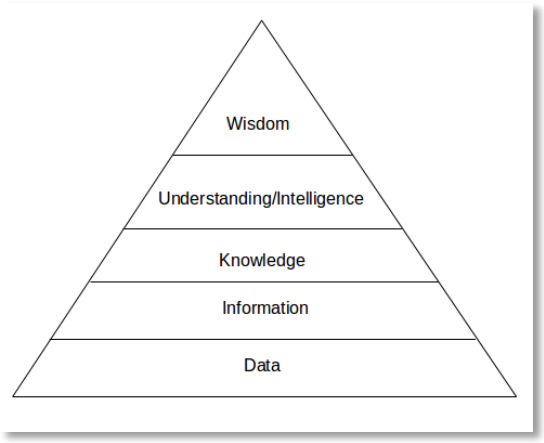


Figure 4.2: Hierarchy of human understanding (Ackoff, 1989)

Since the Ackoff hierarchy of human understanding was identified in Section 2.2.1 as one of the *kernel theories* on which the design of the solution is based and since this section aims to provide working definitions for *Information Visualisation* and *Knowledge Visualisation*, the elements of the hierarchy are outlined as follows:

- *Data. Data* is considered by Ackoff (1989), Keller and Tergan (2005) and Liew (2013) to be symbols that represent objects. According to Wang (2015), data is "an abstract representation of the quantity of real-world entities or abstract objects according to specific quantified scales" (p. 772). Matney, Avanat and Staggers (2016) considers *Data* on its own as having no meaning.
- *Information. Information* is extracted from data (Ackoff, 1989) and is "contained in descriptions, answers to questions that begin with such words as who, what, where, when and how many" (Ackoff, 1989, p. 3). The statement is supported as well by Rioth, Osterman and Warner (2015). According to Keller and Tergan (2005, p. 3), *Information* is "data that has been given meaning through interpretation by way of relational connection and pragmatic context". The notion of *Information* containing meaning is also supported by other authors (Chen & Floridi, 2013; Liew, 2013; Matney *et al*., 2016; Rioth et al., 2015). In addition, Chen and Floridi (2013) state that for *Data* to be considered as *Information*, it should be well-formed, meaningful and truthful. *Information* is categorised by Chen and Floridi (2013) into the following categories and sub-categories:

    1. Environmental information.
    2. Semantic information.
        a. Instructional information.
        b. Factual information.
            i. True information.

ii. Untrue information.

1. Misinformation.

2. Disinformation.

Chen and Floridi's (2013) categorisation of *Information* is of particular value for this research study since it allows for defining the concept of *facts*, which is contained in the title as well as the research question of the research study. A *fact* can therefore be viewed as semantic factual information, which can be either true or untrue. In the case of untrue information, the untruthfulness can be either intentional (disinformation) or unintentional (misinformation).

- *Knowledge.* Ackoff (1989) defines *Knowledge* as “know-how” (p. 4), which is acquired either through instruction or through experience. Liew (2013) supports *Knowledge* being “know-how” (p. 2) and, in addition, postulates that knowledge is also “know-what” and “know-why” (p. 2). *Knowledge* is furthermore considered by Keller and Tergan (2005) to be *Information* resulting from a cognitive process and which has been added to existing *Knowledge*. Wang (2015) defines *Knowledge* as “acquired and comprehended information generated by the brain embodied as a concept” (p. 776). Rioth, Osterman and Warner (Rioth *et al*., 2015) consider *Knowledge* as the result of adding context and content to *Information*.

- *Understanding.* Ackoff (1989) argues that *Understanding* is the ability of a human being to obtain knowledge by him- or herself. In addition, Liew (2013) considers *Understanding* to be a human being’s thought or mental processing abilities. Nurnberger *et al*. (2009) do not consider *Understanding* to be part of the hierarchy of human understanding, but argues that *Understanding* is an increasing ability as one progresses from data to wisdom. Carlisle (2015) highlights it that *Understanding* is typically excluded as a level from most discussions of the hierarchy. Kirsch *et al*. (2015) supports the statement and in addition states that the reason for the exclusion is due to a consensus in the field that the *Understanding* level should be omitted.

- *Wisdom.* According to Ackoff (1989), *Wisdom* is the ability to increase a human being’s own effectiveness. Matney *et al*. (2016) consider *Wisdom* as the use of *Knowledge* in order to handle or explain problems of a human nature. Literature on *Wisdom* is limited according to Liew (2013) and is therefore not often discussed. A possible reason for this is given by Rioth, Osterman and Warner (2015) who postulate that *Wisdom* is difficult to describe and that there are a general disagreement amongst researchers on what *Wisdom* represents.

For the purpose of the study, the Ackoff hierarchy of understanding is considered as one of the *kernel theories* in which the final software solution and model is grounded. It is proposed that the hierarchy can be applied in order to get an understanding of the possible validity of facts in a textual document set. The Huyser collection of Anglo-Boer War testimonies is

considered as *Data*. *Information* exists in the form of facts extracted from the set of testimonies. By visually representing the coherence of the extracted facts, the user gains *Knowledge* on whether the facts are in coherence or not. An *Understanding* of the coherence theory of truth, which considers the coherence of facts in a set as an indication of truth (see Section 1.2.1), allows for the user to make an assumption about the possible validity of the facts. The ability to make the assumption is considered as *Wisdom*.

Chen *et al*. (2009) suggest that Ackoff (1989) considers *Data*, *Information* and *Knowledge* in the perceptual and cognitive space. Furthermore, Chen *et al*. (2009) postulate that *Information* and *Knowledge* are special forms of *Data* since they can be stored as data in a computational system; and therefore, *Data*, *Information* and *Knowledge* should also be considered in the computational space. It is suggested that since a computational software solution is built for the study, the definitions of provided by Chen *et al*. (2009) for *Data*, *Information* and *Knowledge* should also be considered. Chen *et al*. (2009) define, in the computational space, the concepts of *Data*, *Information* and *Knowledge* as follows:

- *Data* is “computerized representations of models and attributes of real or simulated entities (Chen *et al*., 2009, p. 13)”.

- *Information* is “data that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings” (Chen *et al*., 2009, p. 13).

- *Knowledge,* on the other hand, is “data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning or the transcripts of some knowledge acquired by human beings” (Chen *et al*., 2009, p. 13).

Subsequent to the preceding considerations of *Data*, *Information*, *Knowledge*, *Understanding* and *Wisdom*, working definitions particularly of *Information* and *Knowledge* are presented as follows:

In the context of the research study, **Information** is considered to be:

> *The result of extracting meanings from lower-level data, either through a computational process or by means of human transcription.*

Knowledge is considered to be:

> *The result of a cognitive process, either through computer simulation or by means of human knowledge transcription, resulting in a higher level of human understanding.*

Working definitions for *Data*, *Understanding* and *Wisdom* are not provided or further discussed since the aim of this section is only to provide working definitions for *Information Visualisation* and *Knowledge Visualisation*. The previous discussions of *Data*, *Understanding* and *Wisdom* are considered as sufficient to understand that they are different

from *Information* and *Knowledge* and how they contribute in the hierarchy of human understanding.

Having developed a working definition of *Visualisation* in Section 4.1.2.1 and after having provided working definitions of *Information* and *Knowledge* in this section, the next section consequently considers working definitions of *Information Visualisation* and *Knowledge visualisation*.

4.1.2.3 Working definitions of *Information Visualisation* and *Knowledge Visualisation*

Building on the understanding of the terms *Information* and *Knowledge*, working definitions of *Information Visualisation* and *Knowledge Visualisation* are proposed in this section. In pursuing an understanding of what *Information Visualisation* and *Knowledge Visualisation* are, the views of Eppler and Burkhard (2007), Keller and Tergan (2005) and Burkhard (2004) on the difference between *Information Visualisation* and *Knowledge Visualisation* are explored first.

Eppler and Burkhard (2007) state that the difference between *Information Visualisation* and *Knowledge Visualisation* resides in the *object* (what is being visualised), the *purpose* (why is the information/knowledge being visualised) and the *methods* (how is the information/knowledge being visualised) of the visualisation. Keller and Tergan (2005), on the other hand, postulate that the difference between *Information Visualisation* and *Knowledge Visualisation* resides in the relevant field of study in which the visualisation has its origin. Keller and Tergan (2005) accordingly state that *Information Visualisation* has its origin in the field of computer science and *Knowledge Visualisation* in the field of social sciences.

In addition to Eppler and Burkhard (2007) and Keller and Tergan (2005), Burkhard (2004) states that both *Information Visualisation* and *Knowledge Visualisation* exploit the human ability to process visual representations. According to Burkhard (2004), the difference is situated in how the ability is being used.

Burkhard (2004) therefore argues that *Information Visualisation* “aims to explore abstract data and to create new insights” (p. 519), whereas *Knowledge Visualisation* “aims to improve the transfer of knowledge between at least two persons or groups of persons” (p. 519).

Definitions of *Information Visualisation* and *Knowledge Visualisation* are consequently derived by considering the following (see Table 4.1 on the following page):

- the working definition of *Visualisation* derived in Section 4..1.2.1;
- working definitions of *Information* and *Knowledge* derived in Section 4.1.2.2;
- the preceding discussion of the difference between *Information Visualisation* and *Knowledge Visualisation*.

Table 4.1: Deriving Information Visualisation and Knowledge Visualisation

Concept	Definition	Context extracted from the concept for use in derived definitions
Visualisation	A field of study concerned with the transformation of data to visual representations, where the goal is the effective and efficient cognitive processing of data.	The transformation from data (lower-level of knowledge) to visual representation.
Knowledge	The result of a cognitive process, either through computer simulation or by means of human knowledge transcription, resulting in a higher level of human understanding.	The result of a cognitive process, either through computer simulation or by means of human knowledge transcription.
Information	The result of extracting meanings from lower-level data, either through a computational process or by means of human transcription.	The result of extracting meanings from lower-level data, either through a computational process or by means of human transcription.
Difference between Information Visualisation and Knowledge Visialisation	Different in what, why and how. Different in the relevant field of study (computer science or social sciences). Different in how the visual representations are used. Information Visualisation aims to explore abstract data and to create new insights whereas Knowledge Visialisation aims to improve the transfer of knowledge between at least two persons or groups of persons.	Information Visualisation aims to explore abstract data and to create new insights. Knowledge Visialisation aims to improve the transfer of knowledge between at least two persons or groups of persons

A working definition of **Information Visualisation** is consequently derived as:

> *The transformation from lower-level data to visual representation of meanings extracted from the data, where the extraction is either a computational process or a human transcription process; the aim is to explore data and create new insights.*

It is proposed that from this definition a preliminary process for an *Information Visualisation* solution can be laid out as depicted in Figure 4.3 on the following page. The *data source* and *extract meaning* components is considered as part of the *initial* information extraction model discussed in Chapter 3. For the *intermediate* visual representation model discussed in Section 4.2, the *information*, *visual representation* and *insight acquisition* components are considered as relevant.

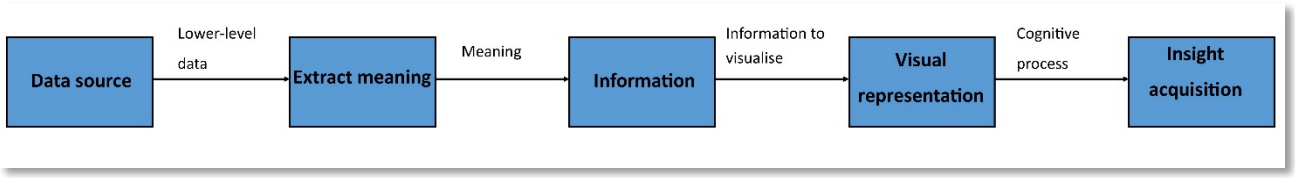


Figure 4.3: Preliminary Information Visualisation process

In the context of this study, the proposed working definition of *Information Visualisation* is then not just concerned with the visualisation representation, but it is also concerned with the extraction of information from lower-level data.

Therefore, an *Information Visualisation* solution consists of two main parts as follows:

- information extraction;
- the information visualisation interface.

Knowledge Visualisation is furthermore defined as:

> *The transformation from knowledge to visual representations of the knowledge, where the knowledge is the result of a cognitive process resulting from either computer simulation or human transcription; the aim is to improve the transfer of knowledge between humans.*

It is suggested by the study that by inspecting the two proposed working definitions of *Information Visualisation* and *Knowledge Visualisation* that they are two distinct fields of study. *Information Visualisation* is considered to be the field of study relevant to this research.

Relevant to *Information Visualisation* is the concept of *Document Visualisation*. The next section briefly discusses Document Visualisation to indicate the relevance.

4.1.2.4 Document visualisation

The concept of *Document Visualisation* is discussed in this section. Various views on *Document Visualisation* exist in the literature as follows:

- *Document Visualisation* is defined by Gan *et al*. (2014, p. 19) as "techniques that transforms textual information such as words, sentences, documents and their relationships into a visual form, enabling users to better understand textual documents and to lessen their mental workload when faced with a substantial quantity of available textual documents". For the purpose of the study, this definition of Gan *et al*. is considered as a working definition of *Document Visualisation.*

In addition, Gan *et al*. (2014) state that *Document Visualisation* allows for the viewing of similarities and connections between documents.

- Lu and Liu (2011) refer to *Text Visualisation* and define it as “the method which uses computer technology to make a graphical show of the specific text” (p. 1), and they furthermore state that the visualisation of text can be based on vocabulary (which includes word clouds and tag clouds), semantics or time series contained in textual documents.

- Gretarsson *et al*. (2012) and Ku *et al*. (2012) claim that the visualisation of textual information contained in documents is more complex than the visualisation of quantitative structured data.

Gan *et al*. (2014) refer to *Document Visualisation* as a sub-category of *Information Visualisation*, and furthermore categorise *Document Visualisation* into three categories as follows:

- visualisations concerned with understanding the content and textual features of a *single document*;

- visualisations concerned with understanding how concepts contained in documents relate *across a number of documents*. This sub-category is of particular interest for the research study. The categories are sub-divided by Gan *et al*. (2014) as follows:

    - visualisation of document themes;

    - visualisation of document core content;

    - visualisation of changes over document versions;

    - visualisation of document relationships;

    - visualisation of document similarities;

- visualisations concerned with comprehensive tasks that are beyond document attributes and are applied to a specific domain.

In this section the concept of *Document Visualisation* as a relevant and sub-category of *Information Visualisation* has been described. It is proposed that the study could therefore, in addition to being labelled as an *Information Visualisation* study, also be considered as a study in *Document Visualisation.* The definition as provided by Gan *et al*. (2014) was indicated as a working definition of *Document Visualisation.*

Having developed working definitions of *Visualisation*, *Knowledge*, *Information*, *Information Visualisation*, *Knowledge Visualisation* and *Document Visualisation* from the literature, the next section summarises these definitions.

4.1.2.5 Summary

The working definitions of *Visualisation*, *Information*, *Knowledge*, *Information Visualisation* and *Knowledge Visualisation* where evaluated through a process of peer review and academic scrutiny by means of an accepted conference paper (Engelbrecht *et al*., 2014) and journal publication (Engelbrecht *et al*., 2015). A summary of the visualisation concepts is given in Table 4.2.

Table 4.2: Summary of working definitions of visualisation concepts

Concept	Working definition
Visualisation	*A field of study concerned with the transformation of data to visual representations, where the goal is the effective and efficient cognitive processing of data.*
Information	*The result of extracting meanings from lower-level data, either through a computational process or by means of human transcription.*
Knowledge	*The result of a cognitive process, either through computer simulation or by means of human knowledge transcription, resulting in a higher level of human understanding.*
Information visualisation	*The transformation from lower-level data to visual representation of meanings extracted from the data, where the extraction is either a computational process or a human transcription process; the aim is to explore data and create new insights.*
Knowledge visualisation	*The transformation from knowledge to visual representations of the knowledge, where the knowledge is the result of a cognitive process resulting from either computer simulation or human transcription; the aim is to improve the transfer of knowledge between humans.*
Document visualisation	*Techniques that "transforms textual information such as words, sentences, documents and their relationships into a visual form, enabling users to better understand textual documents and to lessen their mental workload when faced with a substantial quantity of available textual documents" (Gan et al., 2014, p. 19).*

For the purpose of the research study, *Information Visualisation* is considered to be the visualisation option with which the study concerns itself. The views of Gan *et al*. (2014) on *Document Visualisation* are adopted for the study and as a result *Document Visualisation* is considered as sub-category of *Information Visualisation*.

Having presented working definitions of various visualisation concepts in this section, a set of *Information Visualisation* design principles and guidelines constructed from the literature is presented in the next section.

4.1.3 Information Visualisation design principles and guidelines

A set of principles and guidelines to guide the design of an Information Visualisation interface is presented in this section. According to Ku *et al*. (2012), visualisation principles and guidelines should guide the design of the interface and should not be considered as strict rules. According to Van Biljon and Renaud (2015), defining Information Visualisation principles and guidelines remains a research challenge.

The design principles and guidelines presented in this section serve two purposes:

1. to provide by means of a literature review an answer to the second research objective: What visualisation design principles can be applied to guide the design of an information visualisation interface?

2. in developing a model to visually represent the coherence of facts in a document set, design science research is conducted. As part of the research a specific information visualisation solution is built and instantiated. The *rigour cycle* of design science research furthermore requires that the research should be based on a scientific knowledge base (Hevner, 2007). By constructing a set of Information Visualisation interface design principles and guidelines from available literature, a scientific knowledge base is provided to the design of the Information Visualisation interface.

A review of literature relevant to principles and guidelines to guide the design of an Information Visualisation interface, revealed that design principles and guidelines could grouped according to the following concerns:

- user interactions and interfaces (Section 4.1.3.1);
- additional information required by the user of the visualisation (Section 4.1.3.2);
- the profile of the user (Section 4.1.3.3);
- the Eppler and Burkhard (2007) Knowledge Visualisation framework (Section 4.1.3.4);
- Forsell and Johansson's (2010) heuristic set for Information Visualisation evaluation (Section 4.1.3.5);
- Elmqvist and Fekete's (2010) guidelines for implementing visual hierarchical aggregates (Section 4.1.3.6);
- visualisation shortcomings (Section 4.1.3.7).

In the next seven sections, the listed concerns relating to principles and guidelines on the design of an Information Visualisation interface are discussed.

4.1.3.1 Information Visualisation design principles and guidelines related to user interactions and interfaces

Literature related to design principles and guidelines taking user interactions and interfaces into consideration is discussed in this section. The intended interactions of the user with the visualisation are viewed by Dix (2013), Shneiderman (1996) and Ku *et al*. (2012) as an important consideration with regard to the design of the visualisation. (Liu *et al*., 2014) state that the selection of the user interactions is just as important for the effective understanding

and analsysis of the presented information as the visualisation technique used. Furthermore, Dix (2013) postulates that the design of a visualisation starts with a consideration of the tasks that the user should be able to perform. In addition, Dix (2013) categorises user interaction tasks according to whether the user is a data consumer or a data analyst.

Focus-and-context, details-on-demand and brushing-and-linking user interactions are highlighted by Ku *et al*. (2012) as important user interactions to consider in the design of a visualisation solution. The focus-and-context or fisheye-view user interaction is also discussed by Dix (2013) and Wu, Liu, Liu and Ma (2013). The focus-and-context, details-on-demand and brushing-and-linking user interactions are described as follows:

- The *focus-and-context* or fisheye-view interaction enables the user to focus on specific information while getting an overview of how the information relates to the bigger context (Ku *et al*., 2012).

- The *details-on-demand* interaction enables the user to only view information details when requested, helping to overcome problems with regard to limited display size and cognitive overload by presenting to the user only what he or she needs to see (Ku *et al*., 2012).

- The *brushing-and-linking* interaction allows for the change in one view to dynamically change another view (Ku *et al*., 2012).

In addition to the mentioned user interactions, the research study considers Shneiderman's (1996) visual information seeking mantra, which is referred to by various authors including Keller and Tergan (2005) , Gan *et al*. (2014) and Dragisic, Ivanova, Lambrix, Faria and Jim (2016).

According to Shneiderman (1996), his visual information seeking mantra provides the starting point for designing the interface of a visualisation, by defining seven typical tasks that a user performs when searching for information. The user interaction tasks provided by Shneiderman are listed as follows:

1. Get an *overview*.

2. *Zoom in* on an item.

3. *Filter* out the items not of interest.

4. Get *details-on-demand*.

5. View how the items *relate* to each other.

6. Keep a *history* of the user interactions.

7. Allow the user to *extract sub-collections* of the information.

In addition to considering user interactions, the interface presented to the user can also be taken into consideration for inclusion into a set of design principles and guidelines. Concerns related to the design of user interface of the visualisation are provided by Dix (2013). These concerns are furthermore included in the set of design principles and guidelines. Therefore, in the design of the user interface of the visualisation, the following questions provided by Dix (2013) can be considered:

- What can the user see?
- What does the user need to see?
- What would the user like to see?

The preceding questions can be used as a starting point for designing the interface of a visualisation. Grainger, Mao and Buytaert (2016) similarly state that the design of the interface should be user-centred, thereby considering user characteristics such as what the user needs, wants and is limited by.

Design principles and guidelines for a visualisation interface according to user interaction tasks and to what the user can, need and would like to view were discussed in this section. Visualisation design principles and guidelines related to additional information required by the user of the visualisation are provided in the next section.

4.1.3.2 Information Visualisation design principles and guidelines related to additional information required by the user

The user's need for additional information required to make sense of an Information Visualisation is discussed by various authors including Chen (2005), Chen *et al*. (2009), Gurupur *et al*. (2012), Huang *et al*. (2015) and Grainger *et al*. (2016). Literature related to design principles and guidelines taking additional information required by the user into consideration is therefore discussed in this section. There are three types of additional information, according to Chen (2005) and Chen *et al*. (2009), which the user should have access to in order to make sense of the visualisation and to assist with the insight-acquiring process as follows:

- prior information on how to operate the visualisation;
- prior domain information to assist the user in interpreting the visualisation;
- information about the data itself.

Taking into account that the user should be presented with prior domain information and information about the data itself, it is proposed that a *domain information pipeline, document set information pipeline* and a *document information pipeline* component could be added to the preliminary *Information Visualisation* process depicted in Figure 4.3.

Having access to prior domain information and information about the data itself is considered by various authors including (Gurupur *et al*., 2012; Huang *et al*., 2015) as important for assisting the user to acquire in insight into the data. Gurupur *et al*. (2012) in particular have shown that by providing additional medical domain knowledge (in the form of concept maps) to the users of a health system, the users were able to analyse their own health and receive medical advice.

The need for providing information on how to operate the device is also considered by Grainger *et al*. (2016) and Wu, Lehman and Dunaway (2015). Prior information on how to operate the visualisation is not added as a component of the visualisation process as it is considered by the study as part of the *visual representation* component, whose need only becomes clear in the complexity of the visualisation when the visual interface is designed. As a result of considering the provision of additional information to the user, a proposed *Information Visualisation* process is depicted in Figure 4.4.

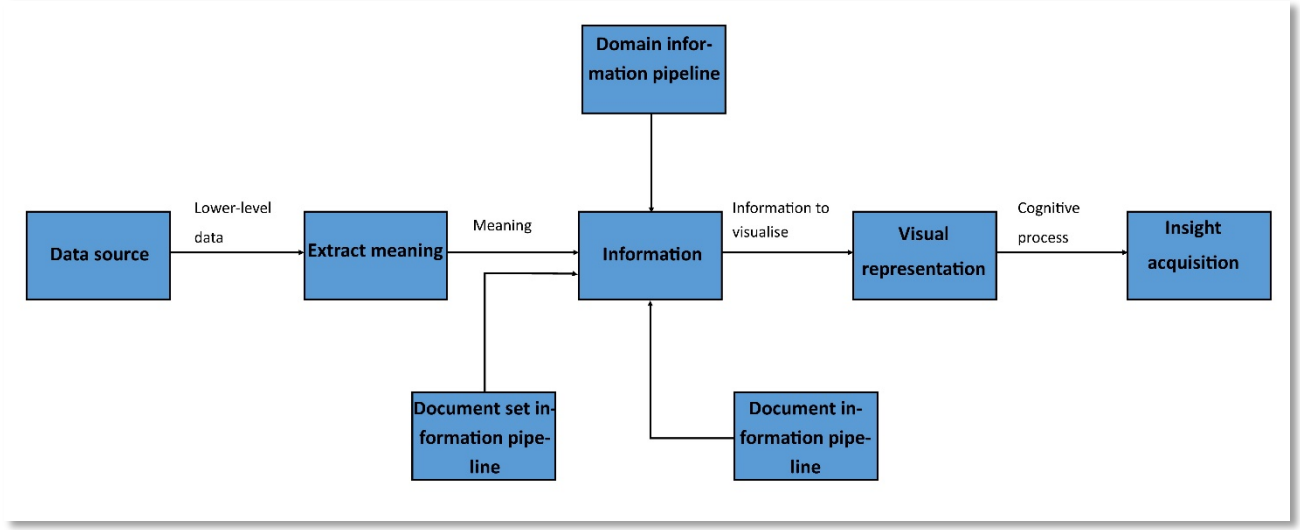


Figure 4.4: Proposed main components of an Information Visualisation solution

Having discussed design principles and guidelines related to additional information required by the user, the next section discusses design principles and guidelines related to the profile of the user.

4.1.3.3 Information Visualisation design principles and guidelines related to the profile of the user

Various authors including Bihanic and Polacsek (2012), Grainger *et al*. (2016), Nafea, Maglaras, Siewe, Smith and Janicke (2016), Dragisic, Ivanova, Lambrix, Faria and Jim (2016) and Capuano, Aniello, Gaeta and Mirande (2015) regard considerations around the profile of the user as of importance during the design of a visualisation. Literature related to design principles and guidelines taking the profile of the user into consideration is therefore discussed in this section. Bihanic and Polacsek (2012) state that providing for the specific profile of users should be considered in the design of a visualisation solution. Furthermore,

Bihanic and Polacsek (2012) state that the visualisation should allow for the restructuring of the visual representation in accordance with requirements defined for the specific user profile. The user profile, together with the semantic data structure, is considered by Bihanic and Polacsek (2012) as the main input into the visualisation design.

In addition other authors also highlight the importance of taking the profile of the user in account when designing an information system, including the following:

- Grainger *et al*. (2016) postulate that the specific user audience should be considered in the design since users have different cognitive abilities, personal traits, domain expertise and cultural, institutional and disciplinary backgrounds.

- Nafea *et al*. (2016) state that users have different learning styles and that system designers should take it into consideration during the design phase. In addition, they suggest that systems could be built in an adaptive manner to accommodate the different learning style profiles.

- User profiles are considered by Dragisic *et al*. (2016) according to the domain expertise, technical expertise and system expertise of the user. These profiles are considered as being related to the additional information required by the user discussed in Section 4.1.3.2.

- According to Capuano *et al*. (2015) different types of users prefer different interaction methods.

- Different types of users are considered by Nafari and Weaver (2015) according to their linguistic knowledge, technical knowledge and domain knowledge. These user types are considered as being related to the additional information required by the user discussed in Section 4.1.3.2.

Having discussed design principles and guidelines related to the profile of the user, the next section discusses design principles and guidelines corresponding to the Eppler and Burkhard (2007) knowledge visualisation framework.

4.1.3.4 Information Visualisation design principles and guidelines corresponding to the Eppler and Burkhard (2007) knowledge visualisation framework

Design principles and guidelines that correspond to the Eppler and Burkhard (2007) knowledge visualisation framework are discussed in this section since they provide typical interface design questions. Other frameworks also exist in literature but they don't focus on the design of the visualisation:

- The framework provided by Grainger *et al*. (2016) for instance focus on the design process, splitting the process into a preparation phase, development phase, visual encoding phase and an analysis and evaluation phase. Similarly the framework provided by McKenna, Mazur, Agutter and Meyer (2014) define the design process

as consisting of tasks namely the understanding of the problem, the generating of solution ideas, the development of a prototype and the deploying of the solution

- The Spiegel framework (Bischof, Dale, & Peterson, 2006) and TULIP framework (Auber *et al*., 2016) each provide a suite of tools and techniques for solving visualisation problems.

Eppler and Burkhard's (2007) knowledge visualisation framework, depicted in Figure 4.5, provides principles and guidelines for designing a visualisation, providing five questions to consider in the design of a knowledge visualisation solution as follows:

1. What is the type of knowledge to be visualised?
2. Why is the knowledge visualised?
3. For whom is the visualisation intended?
4. What is the context in which the visualisation take place?
5. How can the knowledge be presented?

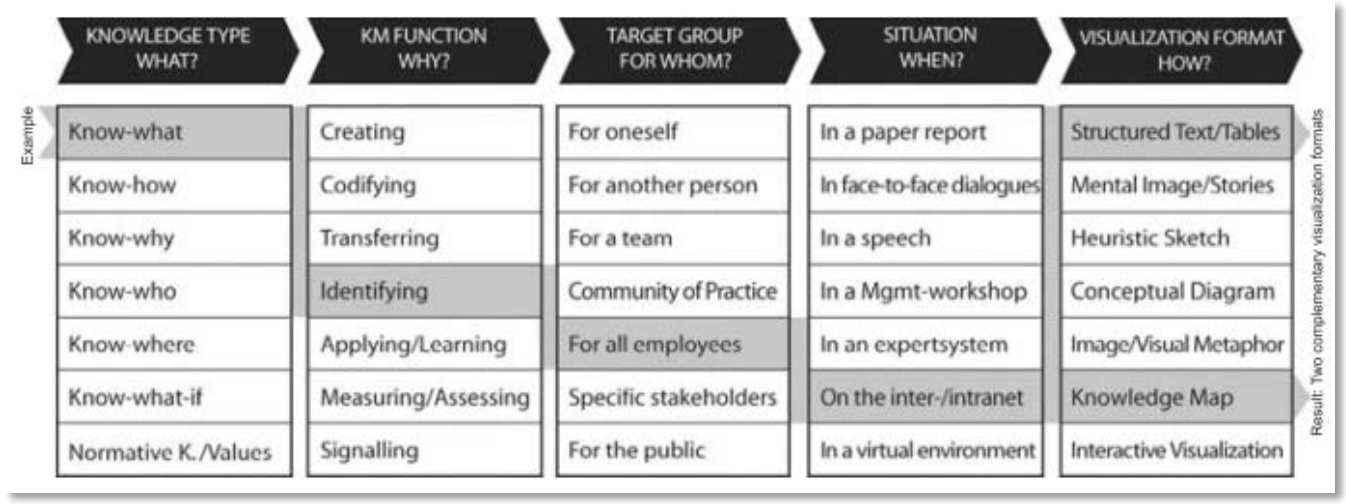


Figure 4.5: The Eppler and Burkhard (2007) knowledge visualisation framework

According to Eppler and Burkard (2007), the difference between Knowledge Visualisation and Information Visualisation is grounded in *what* is visualised, *why* it is visualised and *how* it is visualised. These questions correspond to Questions 1, 2 and 5 of the Eppler and Burkhard knowledge visualisation. Therefore, the same Eppler and Burkhard (2007) knowledge visualisation questions can be considered in the design of both Information

Visualisation and Knowledge Visualisation solutions, however, with expected similar answers to Questions 3 and 4 and dissimilar answers to Questions 1, 2 and 5.

Consequently, the Eppler and Burkhard (2007) Knowledge Visualisation framework can also be applied to Information Visualisation and therefore be incorporated into the set of Information Visualisation design principles and guidelines.

Having discussed the design principles and guidelines related to the Eppler and Burkhard (2007) Knowledge Visualisation framework, the next section considers design principles and guidelines related to the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation.

4.1.3.5 Information Visualisation design principles and guidelines related to the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation

Heuristics evaluation which involves the evaluation of a system against a set of heuristics, is a popular method for evaluating the usability of a system (Hermawati & Lawson, 2016). The study proposes that by considering usability heuristics during the design of the visualisation, future usability deficiencies can be avoided.

A study by Hermawati and Lawson (2016) identified 70 different heuristic evaluation sets, and for each set determined its application domain, how it was created and how it was validated. The study proposed that the heuristic set provided by Nielsen (1994) contained the most general heuristics and was included in most of the sets. The heuristic sets of Zuk, Schleiser, Neumann, Hancock and Carpendale (2006) and Forsell and Johansson (2010) was indicated as being specific to the Information Visualisation domain. According to Hermawati and Lawson (2016) the heuristic set of Zuk *et al*. (2006) was created by an analysis of theories and frameworks of Information Visualisation, whereas the Forsell and Johansson (2010) heuristic set was created by the following process:

1. Identify existing heuristics including those of Shneiderman (1996) and Nielsen (1994).
2. List the heuristics.
3. Gather usability issues in the form of text and screen shots.
4. Working through the usability issues judging how well each heuristic explain each problem.
5. Identify heuristics that identified the highest percentage of usability problems.

It is deducted from the preceding process that the Forsell and Johansson (2010) heuristic evaluation set summarises heuristics from a number of different heuristic sets and is therefore considered be complete. Santos, Ferreira and Dias (2015) consider the Forsell

and Johansson (2010) heuristic as covering a wider set of heuristics than the Nielsen (1994) and Zuk *et al*. (2006) sets.

For the purpose of the research study, the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation is particularly considered for inclusion in the proposed set of design principles and guidelines. The Forsell and Johansson (2010) heuristic set is also considered by other authors including Blascheck, John, Kurzhanls, Koch and Ertl (2016), Santos *et al*. (2015) and Van Biljon and Renaud (2015) as being appropriate for Information Visualisation.

Based on the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation, the following is considered as principles and guidelines for the design of an Information Visualisation interface:

- *Information coding.* The mapping of data elements to visual objects should be enhanced by realistic techniques or additional symbols.
- *Minimal actions.* The use of minimal user actions to accomplish a goal should be considered.
- *Flexibility.* A number of ways to achieve the same goal should be considered.
- *Orientation and help.* The functionality to control the level of details, redo/undo of actions and representing additional information should be considered.
- *Spatial organisation.* The organisation of the visual layout should be considered.
- *Consistency.* Design choices should be consistently applied to the solution.
- *Recognition rather than recall.* The cognitive load on the user should be minimal.
- *Prompting.* The user should be provided with information on alternatives when several actions are available.
- *Remove the extraneous.* Removing distracting information from the visualisation interface should be considered.
- *Data set reduction.* Means to reduce the data set should be considered.

Having discussed adding the Forsell and Johansson (2010) heuristic set for Information Visualisation evaluation to the proposed set of Information Visualisation design principles and guidelines, the next section discusses Information Visualisation design principles and guidelines related to the Elmqvist and Fekete's (2010) guidelines for implementing visual hierarchical aggregates.

4.1.3.6 Information Visualisation design principles and guidelines related to the Elmqvist and Fekete's (2010) guidelines for implementing visual hierarchical aggregates

Design principles and guidelines that relate to the Elmqvist and Fekete (2010) guidelines for implementing visual hierarchical aggregates are discussed in this section. Elmqvist and Fekete (2010) define a *hierarchical aggregate* as a "grouping of the original data items into a hierarchical structure of data aggregates, each representing their children" (p. 440) and *visual aggregates* as "the graphical depictions of data aggregates" (p. 440).

For the purpose of the research study, a *visual hierarchical aggregate* is therefore considered to be the following:

> *Graphical depictions of the grouping of data or information items into a hierarchical structure where the items and their children are depicted.*

Visual hierarchical aggregates can be used in Information Visualisation solutions (the solution proposed in Section 4.2.1 employs the use of a visual hierarchical aggregate) and therefore the Elmqvist and Fekete (2010) guidelines for implementing visual hierarchical aggregates could be included in the proposed set of principles and guidelines. In addition Wu *et al*. (2013) state that the Elmqvist and Fekete (2010) design techniques can be used to simplify visualisations created for large displays. The guidelines for implementing visual hierarchical aggregates are as follows:

- Maintain a visual *entity budget*.
- Aggregates should *visually summarise* the underlying data.
- *Simplicity*. Aggregate visualisations should be clean and simple.
- Aggregates should be *discriminatingly distinguishable* from data items.
- *Fidelity problems* – keep in mind that visual abstractions may be misleading.
- Visual aggregates should be *interpretable*.

This section proposed that the Elmqvist and Fekete (2010) guidelines for implementing visual hierarchical aggregates could be incorporated into the set of Information Visualisation design principles and guidelines. The next section suggests that issues related to the shortcomings of visualisations could be included in a set of Information Visualisation design principles and guidelines.

4.1.3.7 Information Visualisation design principles and guidelines related to the shortcomings of visualisations

The inclusion of issues related to the shortcomings of visualisations into the proposed set of Information Visualisation design principles and guidelines is discussed in this section. The importance of keeping the shortcomings related to visualisations in mind in the design of the visualisation is highlighted by Keller and Tergan (2005).

Because the concepts of *Information* and *Knowledge* are ambiguous (as discussed in Section 4.1.2), shortcomings described in the literature for both *Information Visualisation* and *Knowledge Visualisation* can be considered for inclusion into the set of Information Visualisation design principles and guidelines.

With regard to *Knowledge Visualisation*, Keller and Tergan (2005) consider the following as visualisation shortcomings:

- visualisation shortcomings related to visualisation tools;
- visualisation shortcomings related to the mapping of concepts to representations.

With regard to *Information Visualisation*, Keller and Tergan (2005) consider the following as visualisation shortcomings:

- technical shortcomings related to the representation of information, specifically where large data structures are used;
- shortcomings related to the visualisation of additional information that is required by the user to make sense of the visualisation.

Various other authors also views technical shortcomings (Liu *et al*., 2014; Van Biljon & Renaud, 2015; L. Wu & Hsu, 2013; Yejun Wu *et al*., 2015; Yingcai Wu *et al*., 2013), visualisation tool shortcomings (Wu *et al*., 2013) and shortcomings related to the mapping of concepts to representations (Dragisic *et al*., 2016; Satyanarayan & Heer, 2014; Wu & Hsu, 2013) as important design considerations.

The inclusion of issues related to visualisation shortcomings into the set of Information Visualisation principles and guidelines was discussed in this section. The proposed Information Visualisation design principles and guidelines are consequenlty compiled from the discussions in the preceding section and are summarised in Table 4.3 on the following page.

Table 4.3: Proposed Information Visualisation design principles and guidelines

Design principle/guideline	References
User interaction and interface concerns	
1. Consider providing focus and context user interactions.	(Dix, 2013; Ku *et al*., 2012)
2. Consider providing details-on-demand user interactions.	(Eppler & Burkhard, 2007; Ku *et al*., 2012)
3. Consider providing brushing-and-linking user interactions.	(Ku *et al*., 2012)
4. Consider providing Shneiderman's (1996) user interactions.	(Gan *et al*., 2014; Keller & Tergan, 2005; Shneiderman, 1996)
a. Provide an overview.	(Elmqvist & Fekete, 2010; Keller & Tergan, 2005)
b. Provide zoom functionality.	(Elmqvist & Fekete, 2010; Keller & Tergan, 2005)
c. Provide functionality to filter out items.	(Elmqvist & Fekete, 2010; Keller & Tergan, 2005)
d. Provide functionality to get details-on-demand.	(Elmqvist & Fekete, 2010; Keller & Tergan, 2005; Ku *et al*., 2012)
e. Provide functionality to show how items relate.	(Keller & Tergan, 2005)
f. Provide functionality to keep history of user interactions.	(Dix, 2013; Keller & Tergan, 2005)
g. Provide functionality to extract sub-collections.	(Elmqvist & Fekete, 2010; Keller & Tergan, 2005)
5. Consider what the user can see, needs to see and would like to see.	(Dix, 2013; Forsell & Johansson, 2010; Grainger *et al*., 2016)
Concerns related to additional information required by the user	
6. Additional domain information required by the user should be provided.	(Chen, 2005; Chen *et al*., 2009; Huang *et al*., 2015; Keller & Tergan, 2005)
7. Additional information on how to operate the visualisation device should be provided.	(Chen, 2005; Chen *et al*., 2009; Grainger *et al*., 2016)
8. Information about the data itself should be provided.	(Chen *et al*., 2009; Huang *et al*., 2015)
Concerns related to the Eppler and Burkhard (2007) framework	
9. The type of information to be visualised should be considered.	(Eppler & Burkhard, 2007)
10. The purpose of the visualisation should be considered.	(Eppler & Burkhard, 2007)
11. The user profile should be considered.	(Bihanic & Polacsek, 2012; Eppler & Burkhard, 2007; Grainger *et al*., 2016)
12. The context in which the visualisation will take place should be considered.	(Eppler & Burkhard, 2007)
13. The visualisation format should be considered.	(Eppler & Burkhard, 2007)
Concerns related to the Forsell and Johansson (2010) heuristic evaluation set	
14. The mapping of data elements to visual objects should be enhanced by realistic techniques or additional symbols.	(Forsell & Johansson, 2010; Keller & Tergan, 2005)
15. Consider the use of minimal user actions to accomplish a goal.	(Forsell & Johansson, 2010)
16. Consider providing a number of ways to achieve the same goal.	(Forsell & Johansson, 2010)

Design principle/guideline	References
17. Consider functionality to provide redo/undo of actions.	(Forsell & Johansson, 2010)
18. The organisation of the visual layout should be considered.	(Dix, 2013; Forsell & Johansson, 2010)
19. Design choices should be consistently applied to the solution.	(Forsell & Johansson, 2010)
20. The cognitive load on the user should be minimal.	(Forsell & Johansson, 2010)
21. Provide the user with information on alternatives when several actions are available.	(Forsell & Johansson, 2010)
22. Consider removing distracting information from the visualisation solution.	(Dix, 2013; Forsell & Johansson, 2010; Keller & Tergan, 2005; Ku *et al*., 2012)
23. Consider means to reduce the data set.	(Forsell & Johansson, 2010; Keller & Tergan, 2005; Ku *et al*., 2012)
Concerns related to Elmqvist and Fekete's (2010) visual aggregate implementation guidelines	
24. Maintain a visual entity budget.	(Elmqvist & Fekete, 2010)
25. Visual aggregates should visually summarise the underlying data.	(Dix, 2013; Elmqvist & Fekete, 2010)
26. Visualisations should be clean and simple.	(Elmqvist & Fekete, 2010)
27. Visual aggregates should be discriminatingly distinguishable from data items.	(Elmqvist & Fekete, 2010)
28. Keep in mind that visual abstractions may be misleading.	(Dix, 2013; Elmqvist & Fekete, 2010)
29. Visual aggregates should be interpretable.	(Dix, 2013; Elmqvist & Fekete, 2010)
Concerns related to shortcomings of visualisations	
30. Shortcomings related to the visualisation tools should be considered.	(Keller & Tergan, 2005)
31. Shortcomings related to the mapping of concepts with representations should be considered.	(Forsell & Johansson, 2010; Keller & Tergan, 2005)
32. Shortcomings related to technical problems to represent information should be considered.	(Keller & Tergan, 2005)

Value is added to the proposed design principles and guidelines through a process of peer review and academic scrutiny by means of an accepted conference paper (Engelbrecht *et al*., 2014) and journal publication (Engelbrecht *et al*., 2015).

A set of Information Visualisation design principles and guidelines to guide the interface design of an Information Visualisation solution was presented in this section. The Information Visualisation design principles and guidelines provide an answer to the second research question: *What visualisation design principles could be applied to guide the design of an information visualisation interface*?

Having provided a set of Information Visualisation principles and guidelines, literature with regard to tools and techniques available for implementing the Information Visualisation interface design is discussed in the next section.

4.1.4 Information Visualisation tools and techniques

Literature relevant to Visualisation tools and techniques is discussed in this section. Literature on Visualisation techniques is discussed in Section 4.1.4.1 and in Section 4.1.4.2, tools that could be used to implement the interface of a Visualisation are provided.

4.1.4.1 Literature review of Visualisation techniques

Literature describing Visualisation techniques are identified and discussed in this section with specific reference to Bihanic and Polacsek (2012) and Harger and Crossno (2012). Wang, Perez-Riverol, Hermjakob and Vizca (2015) considers only three main categories namely charts, networks and hierarchies and are therefore not further discussed.

Bihanic and Polacsek (2012) and Harger and Crossno (2012) group Visualisation techniques into families of general techniques as depicted in Table 4.4. These general families of techniques serves as a starting point for the study to get an insight into some of available Visualisation techniques.

Table 4.4: General Visualisation techniques

Family of visualisation techniques	Bihanic and Polacsek (2012) general visualisation techniques	Harger and Crossno (2012) general visualisation techniques
Trees	x	x
Maps/Graphs	x	x
Landscapes		x
Multidimensional views	x	
Geospatial/spatial-temporal views	x	

The families of Visualisation techniques listed in Table 4.4 are briefly discussed as follows:

- *Trees:* Visualisation techniques used to visually organise data (Bihanic & Polacsek, 2012);
- *Maps/Graphs:* Visualisation techniques used to indicate connections between data elements (Bihanic & Polacsek, 2012). The Visualisation presents nodes and edges, where nodes represent data and edges represent the connections between the data (Harger & Crossno, 2012);
- *Landscapes:* Visualisation techniques used to display data side by side in order to indicate space, time and logical relationships (Bihanic & Polacsek, 2012);
- *Multidimensional views:* Visualisation techniques used for data that can be represented as rows and columns of a table (Harger & Crossno, 2012);
- *Geospatial/spatial-temporal views:* Visualisation techniques used to view geographical data (Harger & Crossno, 2012).

The families of Visualisation techniques presented by Harger and Crossno (2012) and Bihanic and Polacsek (2012) focus on general visualisations. Ku *et al*. (2012) however divide the techniques available for *Document Visualisation* into tag/word clouds, document clustering, tree-based graph, circular graph and network (node and link) graph visualisations. For the purpose of the research study, the *Document Visualisation* techniques defined by Ku *et al*. (2012) are considered more relevant to study as they address *Document Visualisation* which was indicated, for the purpose of the study, as a relevant sub-category of *Information Visualisation* (see Section 4.1.2.4).

Each of these *Document Visualisation* techniques defined by Ku *et al*. (2012) are now briefly discussed as follows:

- *Tag/word clouds:* Visualisation techniques employed to visualise the frequency of data contained in textual sources (Devendorf, 2012; Gan *et al*., 2014; Grant *et al*., 2015; Lu & Liu, 2011). According to Emerson *et al*. (2013), tag/word cloud visualisations are simple in their technique and implementation. They further postulate that their potential has not yet been fully explored, and propose that tag/word clouds can become a more powerful Visualisation tool by extending its current functionality. Emerson *et al*. (2013) state that tag/word clouds are useful in performing a variety of user tasks including getting an overall view of data, determining the absence of information from data, studying quantity distributions in data, exploration of co-relationships and relations between elements in data, and lastly, discovering new knowledge in data.

  According to Ku *et al*. (2012), the difference between a tag cloud and a word cloud visualisation resides in the type of frequency being visualised; a tag cloud is a visualisation of theme frequencies and a word cloud is a visualisation of word frequencies in a textual source set. Lu and Liu (2011) also define tag clouds as being the visualisation of theme frequencies in a textual source set.

- *Document clustering:* A Visualisation technique employed to group documents from a document set into clusters of related documents (Ku *et al*., 2012).

- *Tree-based graph:* Ku *et al*. (2012) regard the tree-based graph Visualisation technique as a *Document Visualisation* technique that can be employed in scenarios where hierarchical information contained in textual sources needs to be visually represented.

- *Network (node and link) graph:* A Visualisation technique employed to visualise relationships between entities and topics contained in a textual document set (Devendorf, 2012; Gretarsson *et al.*, 2012; Rusu *et al*., 2009). Rusu *et al*. (2009) additionally employ node and link graphs to visually represent the semantic structure contained in textual documents.

Gretarsson *et al*. (2012) and Devendorf (2012) propose that by adding colour schemes to nodes and links in a graph, additional information about the data itself, for instance, the source of the data, can be visually represented. Furthermore, Gretarsson *et al*. (2012) and Devendorf (2012) employ node and link graphs to visualise similarities between topics in a set of textual documents.

- *Topics* are defined as probability distributions of words in a textual document set (Devendorf, 2012; Gretarsson *et al*., 2012), which are extracted from a textual document set by means of an unsupervised, machine-learning technique known as statistical topic modelling (Devendorf, 2012; Gretarsson *et al*., 2012). Each topic is presented as a set of words that describes a theme (Devendorf, 2012; Gretarsson *et al*., 2012). By visually displaying the similarity between topics contained in each document in the set, similar documents can be visually identified and grouped into document clusters (Gan *et al*., 2014; Gretarsson *et al*., 2012). Gretarsson *et al*. (2012) further state that each individual document can be broken up into sections on which topic modelling and, subsequently, a visualisation of the topics can be applied to discover additional information in the documents.

- *Circular graphs:* A Visualisation technique, which is identified by various authors (Michael Bostock, 2012; Devendorf, 2012; Fernanda *et al*., 2013; Gretarsson *et al*., 2012; Ku *et al*., 2012; Zhao & Lu, 2015) and which uses circular structures to visually represent information.

  The combination of node-and-link graphs in a circular treemap with circular balloon-type layouts is considered by Viégas *et a*l. (2013) as an effective Visualisation technique. An example of combining node-and-link graphs with circular graphs into a circular treemap is provided by Viégas *et al*. (2013) and is depicted in Figure 4.6 on the following page.

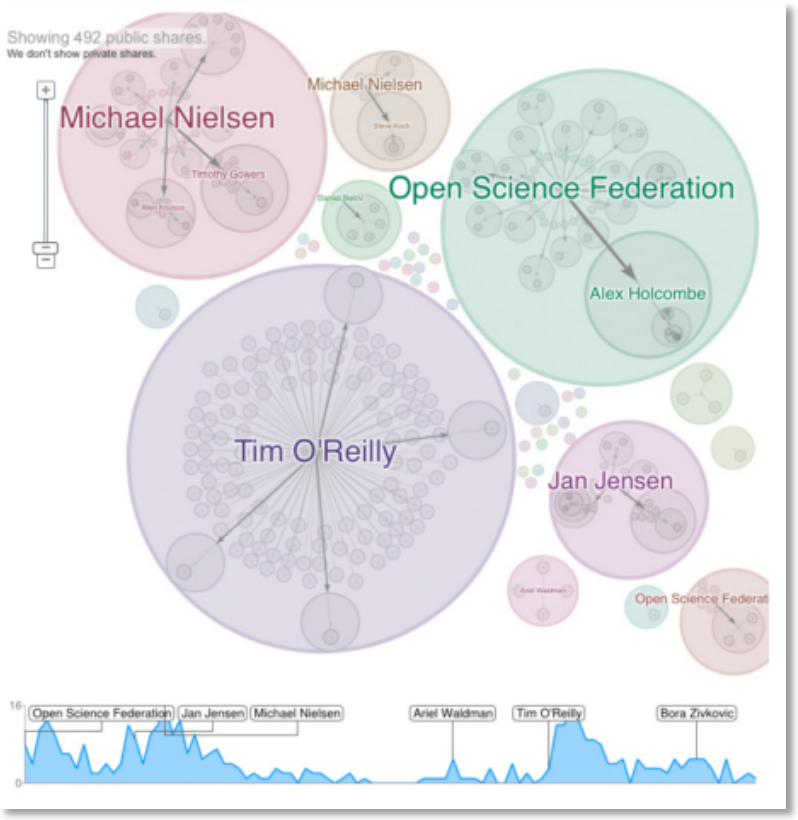


Figure 4.6: Circular treemap example provided by Viégas *et al*. (2013)

For the purpose of the research study, a circular graph technique can be used to visualise the coherence of facts contained in a textual document set for the following reasons:

- It is considered by various authors (Bostock, 2012; Devendorf, 2012; Fernanda *et al*., 2013; Gretarsson *et al*., 2012; Ku *et al*., 2012; Zhao & Lu, 2015) as an effective Visualisation technique.

- On inspection of an example of a circular graph such as the one in Figure 4.6, it is proposed that a circular structure can fit in with the type of information the study aims to visualise, namely an overview of the whole set of documents with an indication of each factual piece of information inside each document. It is therefore proposed that a circle structure can be used to:

    - indicate the document set as a whole by one big circle,

    - consisting of smaller circle each representing a document in the set and

    - representing facts in each document by a third level of the smallest circles.

- It is proposed that fact coherence can be easily indicated on a circular graph by marking all facts (indicated by circles) that are in coherence with the same colour.

- The study follows a pragmatic approach as discussed in Chapter 2. According the philosophical view of pragmatism, truth is based on what works at the time to address the research problem (Creswell, 2014; Gray, 2013). For the purpose of the study, the intend is not necessarily to come up with the best possible Visualisation technique but rather to come up with a suitable one. The choice of using a circular graph for the visual representation is proposed as a possible workable solution. The proposed circular Visualisation technique is discussed in detail in Section 4.2.1.

Having discussed Visualisation techniques identified in the literature, the next section discusses software tools that can be used to implement these techniques.

4.1.4.2 Literature review of Visualisation tools

In the previous section, an overview of a number of Visualisation techniques identified in the literature was provided. In this section, examples of various Visualisation tools that can be used to implement these Visualisation techniques are discussed.

A large number of Visualisation tools, each with a different purpose, are identified by various authors. Table 4.5 contains a list of examples of some of the Visualisation tools identified during the literature review. These list serves as a starting point for the study to get an insight into some of available Visualisation tools and to indicate that a large number of Visualisation tools are available. The list is therefore not considered as comprehensive but rather as a list of example software tools. An evaluation and comparison of these tools are considered as beyond the scope of this research study and therefore only references from the literature are supplied for each Visualisation tool.

Table 4.5: Various visualisation software tools

Visualisation software tool	References
Arc Diagrams	(Gan *et al*., 2014)
ContexTour	(Gan *et al*., 2014)
D3.js	(Bostock, Ogievetsky, & Heer, 2011; Bostock, 2013, 2014; Fernanda *et al*., 2013; Satyanarayan & Heer, 2014; Walny, Huron, & Carpendale, 2015; Wang *et al*., 2015)
DocuBurst	(Gan *et al*., 2014; Rusu et al., 2009)
INSPIRE	(Gan *et al*., 2014; Ku *et al*., 2012; Lu & Liu, 2011; Mooranian *et al*., 2011)
Graphviz	(Auber *et al*., 2016; Recupero *et al*., 2015; Wang *et al*., 2015)
Jigsaw	(Gan *et al*., 2014; Ku *et al*., 2012)
JUNG	(Harger & Crossno, 2012; O'Madadhain *et al*., 2005; Wang *et al*., 2015)
PivotPaths	(Gan *et al*., 2014)
Prefuse	(Heer, Card, & Landay, 2005; O'Madadhain *et al*., 2005; Rusu *et al*., 2009)
Semantic Graphs	(Gan *et al*., 2014)

Visualisation software tool	References
TableLens	(Chi, 2000; Dix, 2013)
Tag Cloud	(Gan *et al*., 2014; Lu & Liu, 2011)
TAGGLE	(Emerson *et al*., 2013)
TIARA	(Gan *et al*., 2014; Liu *et al*., 2014)
TextArc	(Gan *et al*., 2014)
ThemeRiver	(Devendorf, 2012; Gan *et al*., 2014; Liu *et al*., 2014; Lu & Liu, 2011)
Topic Island	(Gan *et al*., 2014)
TopicNets	(Devendorf, 2012; Gan *et al*., 2014; Gretarsson *et al*., 2012)
Wordle	(Emerson *et al*., 2013; Gan *et al*., 2014)
WordTree	(Gan *et al*., 2014; Liu *et al*., 2014; Lu & Liu, 2011; Mooranian *et al*., 2011)

For the purpose of the study, software tools developed in the Java programming language and making use of web application technologies was considered from the list of examples in Table 4.5. The consideration was motivated by the researcher's own professional skillset which includes Java programming and web application development. Building and instantiating a solution using the Java programming language and web application technologies could therefore be a trivial task for the researcher.

From the identified Visualisation tools listed in Table 4.5, a pre-selected number of tools (according to their use of the Java programming language and other web application technologies) namely TAGGLE, Prefuse, JUNG and D3.js, are briefly discussed as follows:

- *TAGGLE.* According to Emerson *et al*. (2013), the software is developed in the Java programming language and stores its data in Extensible Markup Language (XML) files. The tool allows for tag/word clouds to be saved in a variety of formats and, in addition, serves as a platform for research on layout algorithms by allowing for researchers to plug in their own layout algorithms (Emerson *et al*., 2013). Emerson *et al*. (2013) suggest that a useful feature of TAGGLE is the ability to allow for the specification of relationships between tags/words in a tag/word cloud representation.

- *Prefuse.* The Prefuse visualisation toolkit is, as in the case of TAGGLE, developed in the Java programming language (Heer *et al*., 2005; O'Madadhain *et al*., 2005; Rusu *et al*., 2009). The toolkit can be used for a number of visualisations including semantic graph visualisations (Rusu *et al*., 2009). According to Heer *et al*. (2005), Prefuse can be used to simplify the programmatic creation of 2D visualisations, of both structured and unstructured data, which include graph, tree and scatter plot visualisations.

- *JUNG.* A software library, which like TAGGLE and Prefuse is also written in the Java programming language (O'Madadhain *et al*., 2005). The software library is used by software developers to create visualisations of data that can be visually presented as a network graph (O'Madadhain *et al*., 2005). Some of the major features of JUNG, according to O'Madadhain *et al*. (2005), include the support for a variety of

visualisations of entities and the relationships between them, mechanisms to annotate the graphs, the visualisation of entities and their relationships with metadata, implementations of a number of graph-related algorithms, rendering and layout algorithms to produce visualisations of network data and filtering methods to extract subgraphs of network graphs. O'Madadhain *et al*. (2005) postulate that JUNG's separation of its data structures from its visualisation methods makes it possible to use it with other Java visualisation libraries such as Prefuse.

- *D3.js.* A Visualisation framework, which is developed in the Javascript language, combining Hypertext Markup Language (HTML), Scalable Vector Graphics (SVG) and cascading style sheets (*CSS)* technologies to produce a number of visualisations (Bostock *et al*., 2011; Bostock, 2013a, 2014).
  The visualisations made possible by D3.js include basic visualisations such as bar charts and scatter plot charts as well as advanced visualisations such as zoomable circular graphs and dendograms (Bostock, 2014). The data for a visualisation is bound to a document object model (DOM) allowing for a number of data-driven transformations which can be used to produce a number of visualisation techniques (Bostock, 2013a). The JavaScript Object Notation (JSON) data format is used to define the input data (Bostock, 2012).

For the purpose of the research study, D3.js is selected as Visualisation tool for developing a visual representation of the coherence of facts in a document set:

- On inspection of D3.js code examples it was discovered that the *zoomable circle-packing* D3.js code example provided by Bostock (2012) could be a starting point for implementing the interface of the visualisation since it can possibly provide different levels of circular graphs within each other. This can allow for the three level circular structure discussed in Section 4.1.4.1 with the ability to zoom in and out of the different levels; providing an interactive visualisation. The tool provides an easily modifiable example to visually represent the following:
  - the document set as a whole by one big circle,
  - smaller circles each representing a document in the set and
  - a third level of the smallest circles each representing a fact in a document.
- The researcher's working skillset in terms of the technologies (HTML, SVG, *CSS* and JSON) employed by D3.js makes it an effective and relative easy to use tool.
- The researcher's familiarity with the tool from previous work experience convinced the researcher that it is an easy to use, effective and well supported tool.
- The study follows a pragmatic approach as discussed in Chapter 2. According the philosophical view of pragmatism, truth is based on what works at the time to address the research problem (Creswell, 2014; Gray, 2013). The intend of the

study is not necessarily to come up with the best possible Visualisation tool, but rather to come up with a suitable one. D3.js is therefore proposed as a tool suitable for implementing the visual representation of the coherence of facts in a document set.

Various visualisation techniques as well as visualisation tools to implement the techniques were discussed in this section of the literature review. The purpose was not to investigate and compare the visualisation tools and techniques, but to rather give examples of some the tools and techniques. The examples were used as starting point to get an insight into possible tools and techniques for implementing a visual representation of the coherence of facts in a document set.

4.1.5 Summary

Concluding remarks on literature relevant to the visualisation of information coherence are provided in this section. The discussion in this section started with an investigation into work similar to that of the study, thereby identifying the work of Gan *et al*. (2014), Ku *et al*. (2012), Gretarsson *et al*. (2012), Devendorf (2012) and Luo *et al*. (2012).

Literature on defining various working definitions of *Visualisation* concepts in order to provide subject background and to justify the development of an *Information Visualisation* software tool as a suitable insight-acquiring solution was considered next. In order to ensure research rigour and to answer research question 2 defined in Section 1.2.3, a set of Information Visualisation design principles and guidelines (listed in Table 4.3) was compiled from available and relevant literature. The review of literature on Visualisation concepts and design principles and guidelines, lead to the proposed main components of an Information Visualisation process as depicted in Figure 4.4.

Various Information Visualisation methods and tools was considered next. As a result it was decided with motivations that a circular graph technique, implemented using the D3.js software package, would be used in the visual representation.

Using the possible identified main components, methods and tools as input, the next section discuss the instantiation of the specific visual representation solution. During the process of instantiating the solution an *intermediate* visual representation model is developed from the instantiation.

4.2 A MODEL FOR THE VISUAL REPRESENTATION OF THE COHERENCE OF INFORMATION IN A RECORD SET

A proposed *intermediate* model to visually represent the coherence of facts in a document set is presented in this chapter. The model is developed by instantiating the possible components identified in Figure 4.4; excluding the *data source* and *extract meaning* components since they are already included in the initial information extraction model (see

Section 3.3). During the instantiation of the solution the components of the model are refined and sub-components identified. This chapter discusses the design (Section 4.2.1) and building and instantiation (Section 4.2.2) of the specific visual representation software solution. The development of an *intermediate* model for the visual representation of the coherence of facts is afterwards informed (Section 4.3) by the instantiation of the specific software solution.

4.2.1 Proposed design to visually represent the coherence of information in a record set

This section provides a proposed design for the visual representation of information to assist the user in acquiring insight into the coherence of information in a record set. The design of the visualisation artifact consists of the user interface design as well as the design of the database schema to support the view.

4.2.1.1 Design of the user interface

The design of the user interface to assist the user in acquiring insight into the coherence of facts in a document set is discussed in this section. The interactions between each visualisation view presented to the user are described and the design of each view is discussed. In addition, the design of the domain information management and document management pages is discussed. The section concludes with a discussion (Section 4.2.1.1.5) of how the information visualisation principles and guidelines presented was applied to the design of the interface.

Shneiderman (1996) provides seven typical tasks that a user performs when searching for information. These Shneiderman tasks are referenced by various literature sources including Keller and Tergan (2005) and Gan *et al*. (2014) and is therefore considered as appropriate for the research study. Shneiderman's information-seeking tasks are discussed in Section 2.1 as one of the *kernel theories* on which the design is based. Based on Shneiderman's (1996) information-seeking tasks such as that a user should be able to get an *overall view*, *zoom into* specific items and get *details on demand*, the visualisation design presented by this research study in essence consists of three different views (depicted in Figure 4.7 on the following page) as follows:

1. an *overall view*;
2. a *zoom-into-document* view;
3. a *zoom-into-fact* view.

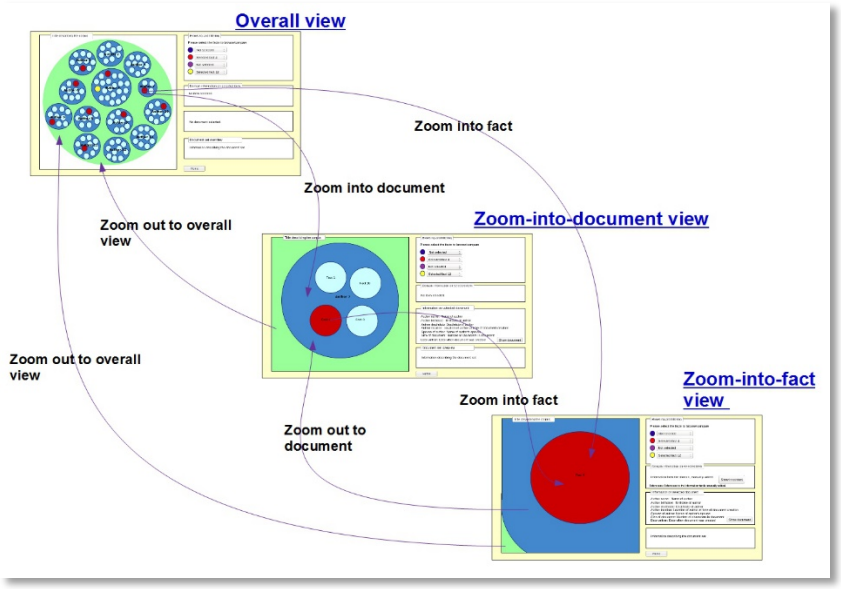


Figure 4.7: Design of the interaction between the visualisation views

The *overall view* represents the record set as a whole, displaying each document and each fact in the set. The details of documents and facts are not displayed on this view. By displaying this view to the user, the user is enabled to get insight into the domain described by the documents. This allows the user to put the facts into an overall context.

The *zoom-into-document view* displays details of a selected document in the record set. Details of facts are not displayed on this view; only the names of each contained fact. The display of this view allows the user to get insight into which facts are being described in the document, thereby giving the user an overall insight into the content of the document. The details of the author of the document allow for the user to make an assumption on the reliability of the document content by considering the age of the author, age at the writing of the document, etc.

The *zoom-into-fact view* displays details of a selected fact contained in a selected document. The details include existing domain information outside the document set, giving additional information supporting or disproving the fact. In addition, information on the document in which the selected fact is contained is also displayed.

The interactions between views of the visualisation are depicted in Figure 4.7. Each view represents a specific level of information, namely the *record set* as a whole, *documents in the set* and *facts* contained in the documents. Furthermore, each view contains a different level of details as described below.

The following interactions can be executed between the three different page views:

- From the *overall view* page, the user can *zoom into* either a document or a fact by clicking on a document or fact-representing circle. Choosing the zoom into a

document presents the user with the *zoom-into-document view* page and choosing the zoom into a fact presents the user with the *zoom-into-fact view* page.

- From the *zoom-into-document view* page, the user can either *zoom into* a fact by clicking on a fact-representing circle or can *zoom out* to the *overall view* page by clicking on the area representing the record set as a whole. The user can therefore either choose to view the *overall view* page or the *zoom-into-fact view* page.

- From the *zoom-into-fact view* page, the user can either zoom out to the *overall view* page or to the *zoom-into-document view* page by either clicking on the document-representing area or the area that represents the record set as a whole.

The user of the interface is equipped by the views of the interface and the interactions between them to acquire insight into the coherence of the facts in the document. Furthermore, multiple occurrences of a selected fact in a document set are an indication whether the fact is in coherence in the set. The coherence of a fact in the document set is an indication of the possible truth value of the selected fact. This proposition is based on the following two theories:

- The coherence theory of truth. For the purpose of the research study, the coherence theory of truth is applied as follows: *The truth of any fact consists in the coherent occurrences of the fact in a document set*.

- The coherence of a proposition with other propositions is an indication of an "increase" in the probability that some truth is contained by the proposition. The proposition of Roche (2013) is, for the purpose of the study, applied as follows: *The coherence of a fact with other occurrences of the fact in a document set is an indication of an "increase" in the probability that some truth is contained by the fact.*

The following two scenarios is used as examples to indicate how the designed interface can help the user to make an assumption about the coherence of a selected fact in the document set.

Scenario 1: *The selected fact is coherently identified in the occurrences of the fact in the document set.*

The user selects a fact (Fact 4) from a dropdown list in order to investigate whether the fact is in coherence with the facts across the documents. Upon selection of the fact, each occurrence of the fact across the documents is highlighted. If, for instance, Fact 4 occurs in 80% of the documents, the user will be able to see 80% of the documents having Fact 4 highlighted in the document fact-representing circles. The user is therefore empowered to make an assumption on the coherence of the fact in the document set. In this case, it might be assumed that Fact 4 is coherently identified in the documents of the document set.

In order to allow the user to make a further assumption on whether Fact 4 contains some truth, researched information from the domain pertaining to Fact 4 is also represented upon zooming into an occurrence of the fact. In the case of Fact 4, it might be (for the purpose of the depicted scenario) that researched information from the domain supports the fact.

In addition, upon zooming into an occurrence of Fact 4, information about the document in which the fact occurs is provided. The information is provided to assist the user in making an assumption about the reliability of the source. The assumption is based on a consideration of the author in terms of the age of the author, age of the author during writing of the document, etc. In the case of Fact 4, it might be (for the purpose of the depicted scenario) that the authors of most of the documents containing the occurrences of Fact 4 were between 30 and 40 years old during the writing of the document, leading to an assumption that the testimonies of the authors might be reliable.

By considering that Fact 4 is coherently identified in the document set, researched domain information supports Fact 4 and the authors of the document containing occurrences of Fact 4 are reliable; there is a possible truth value in Fact 4.

Scenario 2: *The selected fact is not coherently identified in the occurrences of the fact in the document set.*

The user selects a fact (Fact 5) from a dropdown list in order to investigate whether the fact is in coherence with the facts across the documents. Upon selection of the fact, each occurrence of the fact across the documents is highlighted. If, for instance, Fact 5 occurs in 1% of the documents, the user will be able to see 1% of the documents having Fact 5 highlighted in the document fact-representing circles. The user is therefore empowered to make an assumption on the coherence of the fact in the document set. In this case, it might be assumed that Fact 5 is not in coherence in the document set.

In order to allow the user to make a further assumption on whether Fact 5 contains some truth, researched information from the domain pertaining to Fact 5 is also represented upon zooming into an occurrence of the fact. In the case of Fact 5, it might be (for the purpose of the depicted scenario) that researched information from the domain does not support the fact.

In addition, upon zooming into an occurrence of Fact 5, information about the document in which the fact occurs is provided. In the case of Fact 5, it might be (for the purpose of the depicted scenario) that the authors of most of the documents containing the occurrences of Fact 5 were over 80 years of age during the writing of

the document, leading to an assumption that the testimonies of the authors might not be reliable due to memory loss.

By considering that Fact 5 is not in coherence in the document set, researched domain information does not support Fact 5 and the authors of the documents containing occurrences of Fact 5 are unreliable, there is a possibility that Fact 5 does not contain any truth value.

Two scenarios have been used to give examples of how the designed interface could be used to support the user to make an assumption about the truth value of selected facts. The assumption can be based on the coherence of the fact in the document set, researched domain information and information about the authors of the documents in which the facts occur. A multitude of other scenarios is possible, each requiring the cognitive abilities of the user to acquire insight into the truth value of the selected facts, ultimately allowing the user to make an assumption on the validity of the document set as a whole.

Having discussed the interactions between each view of the visualisation and how the views can be used to make assumptions about the truth values of selected facts, the next sections discuss the design of each view.

4.2.1.1.1 The overall view

The *overall view* is discussed in detail in this section. The purpose of the view is to present the document set as a whole to the user. From this view, the user can select to view the coherence of facts in the document set, zoom into documents or zoom into facts.

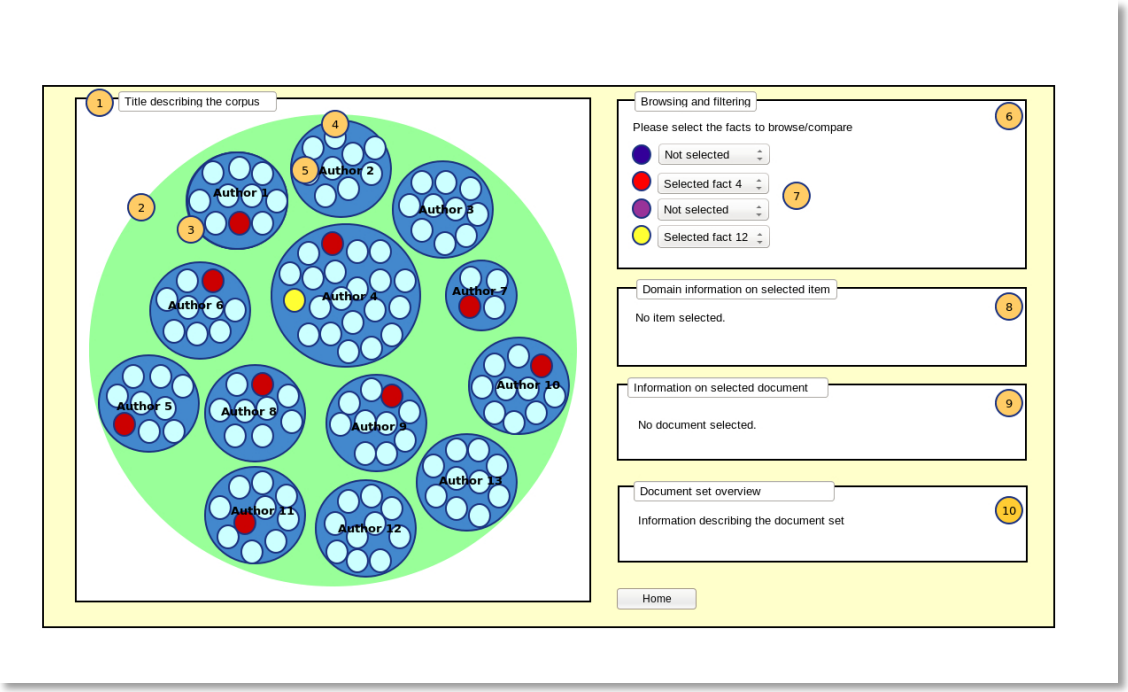


Figure 4.8: Design of the overall view

The design of the *overall view* page is depicted in Figure 4.8 on the previous page. The design is discussed next by referring to each bullet in the figure. The left-hand pane is discussed first, then the purpose of each pane on the right-hand side from top to bottom as follows:

1. A title that describes the record set is presented to the user, indicating the *Corpus section*.

2. The record set as a whole is represented by a large circle.

3. The record set contains documents, which are represented by smaller circles inside the large circle. Each smaller circle represents a document in the set. By clicking on any of the smaller circles, the user is presented with a zoomed-in view of the document as depicted in Figure 4.9.

4. Each document contains facts, which are represented by small circles. Clicking on any of the fact circles presents the user with a zoomed-in view of the fact as depicted in Figure 4.10.

5. For every document, the name of its author is displayed to the user.

6. The *Browsing and filtering* section presents the user with a list of the facts contained in the record set, which can be selected for visually representing its coherence in the set. Upon selecting a fact in one of the dropdowns, all occurrences of the fact in the documents are highlighted; therefore, indicating its coherence in the document set. Each fact can only be selected once in a dropdown; the fact is not available for selection in the other dropdowns. Only four facts can be selected for visually representing the coherence of the facts in order to keep the cognitive load of the user to a minimum.

7. Each fact contained in the document set is listed in a separate dropdown list. Multiple dropdowns are provided in order to allow the user to compare the different facts in terms of coherence and contrast. Selecting a fact in the list automatically updates the fact-representing circles. The colour of each circle representing an occurrence of the selected fact is changed in order to indicate in which documents the fact occurs.

   In the example provided in Figure 4.8, the user has selected Fact 4 in the dropdown list. All occurrences of Fact 4 are automatically indicated by changing the colour of each circle representing Fact 4 to the colour (red) as indicated next to the dropdown list. The coherence of Fact 4 in the document set is therefore visually presented to the user.

   By highlighting each occurrence of a fact in the record set, the coherence of the fact among the documents is visually represented. Furthermore, the visual representation will allow the user to achieve insight into the coherence of the facts contained in the record set.

8. The *Domain information on selected item* section is initially empty and is only populated (see Figure 4.10) when the user clicks on a fact-representing circle. The information displayed contains domain information from domain sources such as academic articles, books or any other published work that can support or reject the selected fact. The information displayed in this section is entered into the system by the user by means of a form as depicted in Figure 4.13.

9. The *Information on selected document* section containing information about a document is initially empty and is only populated (see Figure 4.9) when the user clicks on a document-representing circle. Information about the document itself, including the information on the author and information on the size of the document, is displayed to the user. The information displayed in this section is entered into the system by the user by means of a form as depicted in Figure 4.15.

10. The *Document set overview* section contains a description of the document set in order to give the user background information on the source of the document.

Having described the design of the *overall view*, the next section presents the design of the *zoom-into-document view*.

4.2.1.1.2 The zoom-into-document view

In this section, the design of the *zoom-into-document view* is discussed in detail. The purpose of the view is to present the view of a specific document. Each fact contained in the document is also presented. If any of the facts selected in the dropdown lists is contained in the document, then the fact is highlighted in the document view. From this view, the user can select to zoom into a fact or zoom back to the overall view.

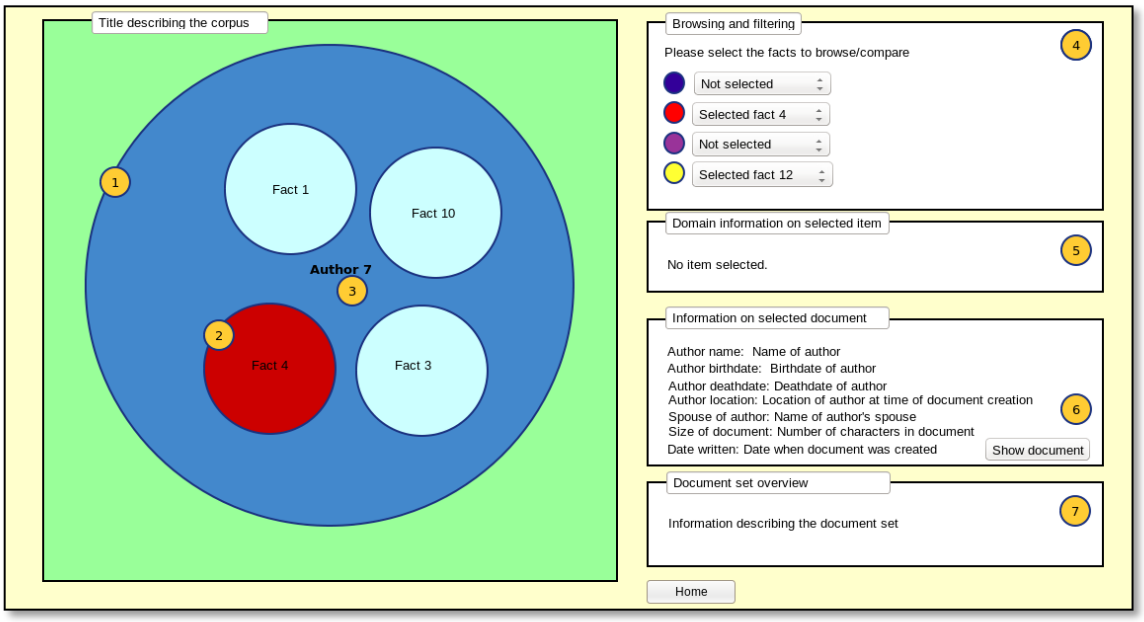


Figure 4.9: Design of the zoom-into-document view

The design of the *zoom-into-document view* page is depicted in Figure 4.9. In this view, the facts contained in one single document are depicted. The design is discussed next by referring to each bullet in the figure as follows:

1. The document is represented by a large circle.

2. Each fact contained in the document is represented by smaller circles. The colour of the circles indicates whether the fact corresponds to any of the facts selected in the *Browsing and filtering* section.

3. The name of the author of the document is presented to the user.

4. The *Browsing and filtering* section presents the user with a list of the facts contained in the record set that can be selected for visually representing its coherence in the set.

5. The *Domain information on selected item* section contains information from the domain describing the fact being zoomed into. The section is empty on this page view because the user has not yet selected a fact to zoom into.

6. The *Information on selected document* section contains information about the selected document. Information about the document itself, including the information on the author and information about the size of the document, is displayed to the user. The information displayed in this section is entered into the system by the user by means of a form as depicted in Figure 4.15.

7. The document set overview section contains a description of the document set in order to give the user background information on the source of the document.

Having described the design of the *zoom-into-document view*, the next section presents the design of the *zoom-into-fact* page.

4.2.1.1.3 The zoom-into-fact view

In this section, the design of the *zoom-into-document view* is discussed in detail. The purpose of the visual interface is to present the view of a specific selected fact in a document. If a fact was previously selected in any of the dropdown lists, then the fact is highlighted in the fact view. From this view, the user can select to either zoom back to the document view or to the overall view.

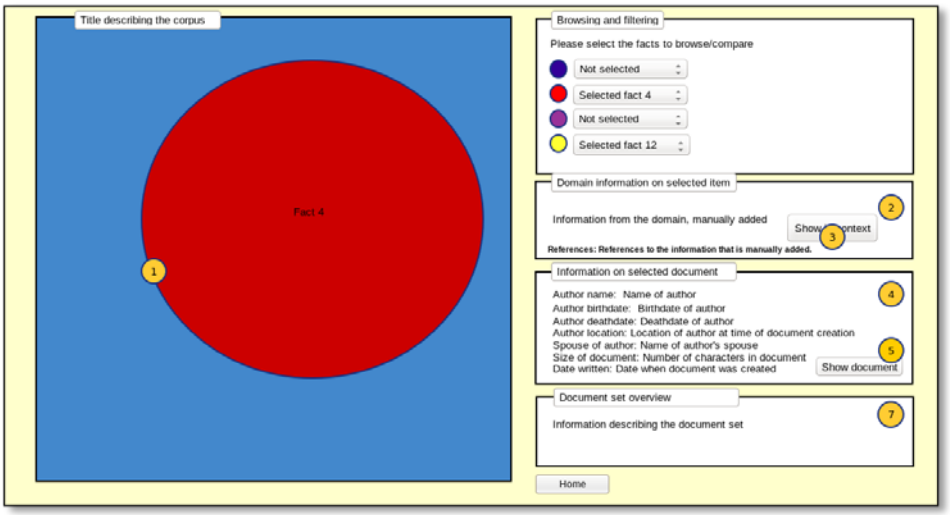


Figure 4.10: Design of the zoom-into-fact view

In this view, one single fact is depicted. The design of the *zoom-into-fact view* page is presented in Figure 4.10. The design is discussed next by referring to each bullet in the figure as follows:

1. The zoomed-into fact is visually represented by a large circle.

2. This view is displayed as a result of the user selecting to zoom into a fact. The *Domain information section* therefore contains information from the domain, describing the selected fact. The information displayed in this section is entered into the system by the user by means of a form as depicted in Figure 4.13.

3. The user can choose to see the context in which the selected fact appears in the document by clicking on the *Show in context* button. By clicking on the button, a dialog similar to the one depicted in Figure 4.11 on the following page is presented to the

user. The appearance of the fact in the transcribed and translated document is highlighted.

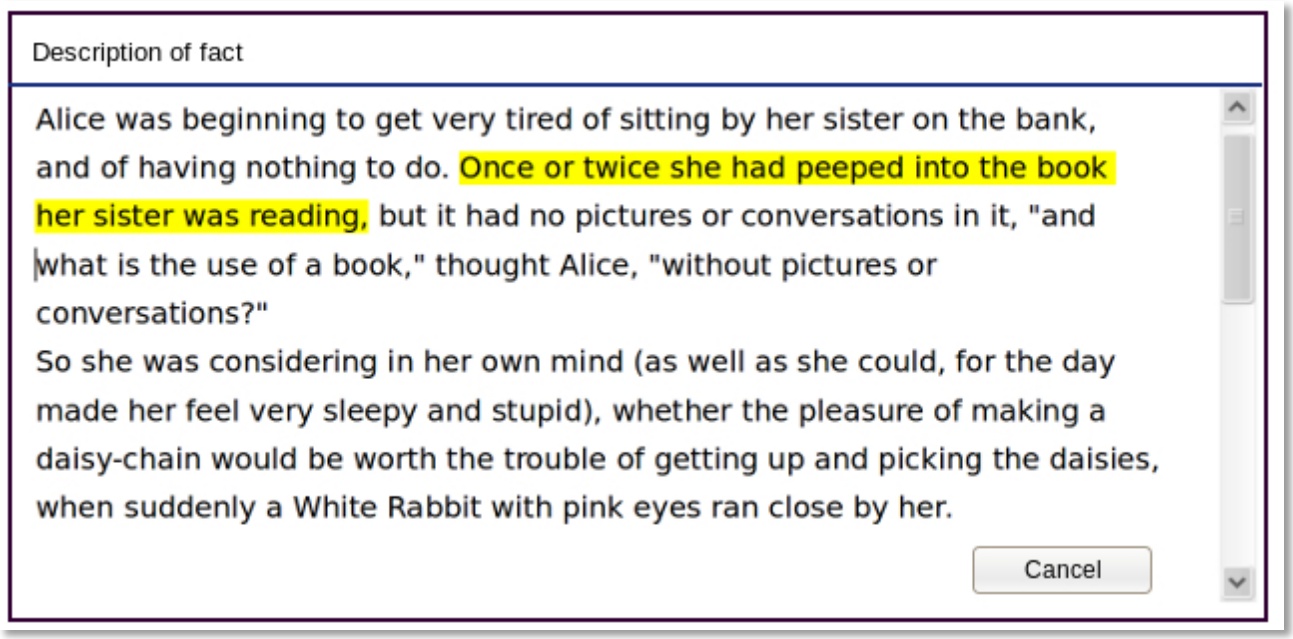


Figure 4.11: Displaying an occurrence of a fact in context

4. Zooming into a fact automatically selects the document that contains the occurrence of the fact. Information about the document itself, including the information on the author and information about the size of the document, is displayed to the user. The information displayed in this section is entered into the system by the user by means of a form as depicted in Figure 4.15.

5. The transcribed and translated document is presented to the user by clicking on the *Show document* button, displaying a dialog similar to the dialog depicted in Figure 4.12 on the following page.

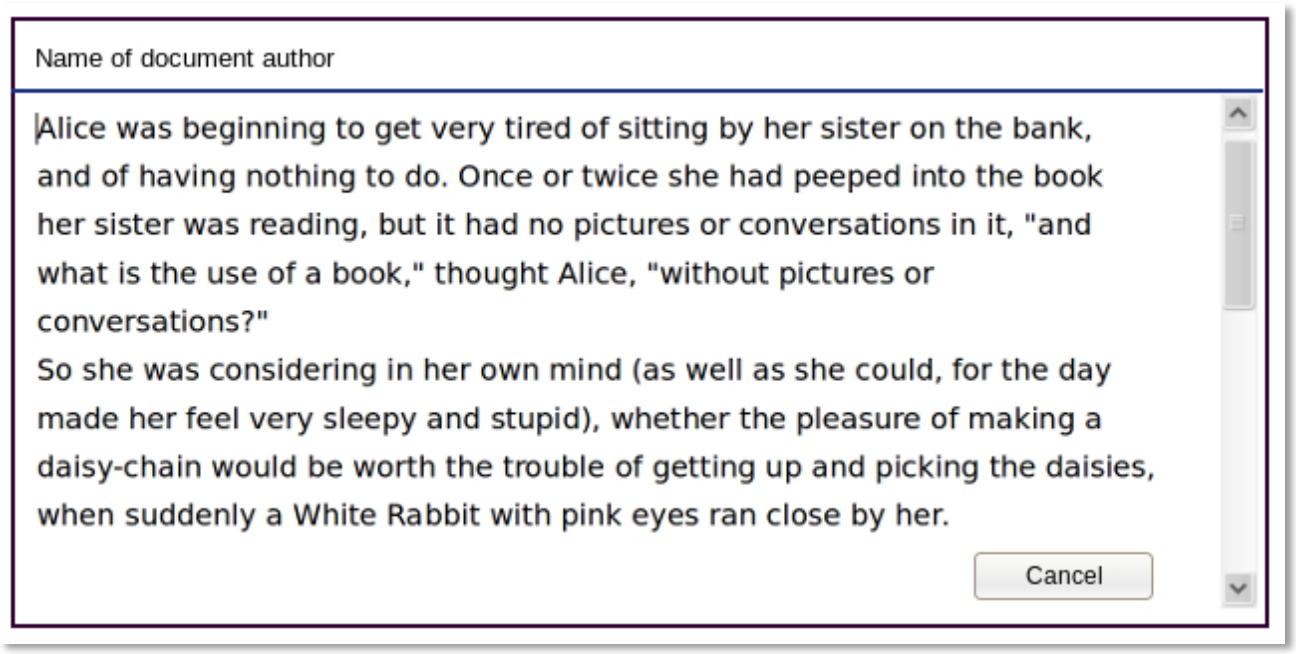


Figure 4.12: Displaying the contents of a document

6. The *Document set* overview section contains a description of the document set in order to give the user background information on the source of the document.

Having discussed the design of the pages that constitute the visual representation of the coherence of facts in a record set, the next section describes the design of the domain information management and document management pages that support the visualisation.

4.2.1.1.4 The domain information management and document management pages

The design of the *domain information management* and *document management* pages that support the *overall view*, *zoom-into-document view* and *zoom-into-fact view* is discussed in this section.

Additional information should be supplied to the user of the visualisation. The additional information can support the user to better understand the information being visualised and can assist the user to acquire insight into the information being visualised. The following additional information is provided to the user of the visualisation:

1. *Domain information* describes existing information gathered from the domain being visualised. The source of the domain information includes sources such as academic articles, books or any other published work that can support or reject the selected fact. By including the domain information in the visualisation, the design adheres to the following design guideline:

   Additional domain information required by the user should be provided.

The domain information and references to a specific fact are entered into the system by means of a manual process. The manual process requires that the user must carry out domain research on a fact and that the user subsequently enter the researched domain information into the system by completing the form depicted in Figure 4.13.

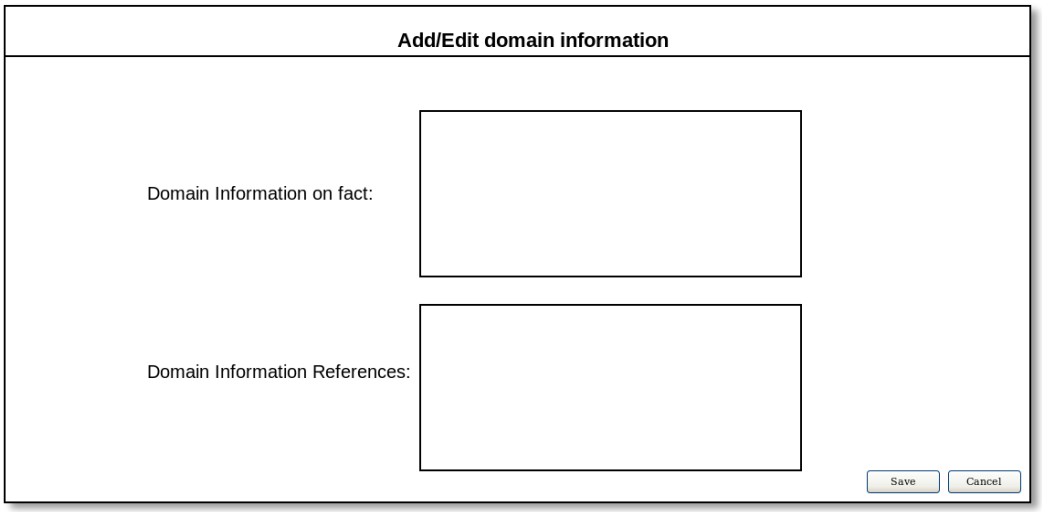


Figure 4.13: Add/Edit domain information page design

All the facts contained in the textual records are listed and displayed to the user as depicted in Figure 4.14. The user can enter domain information on a fact by clicking on the *Add* or *Edit* icon and then subsequently entering the information in the form provided in Figure 4.13.

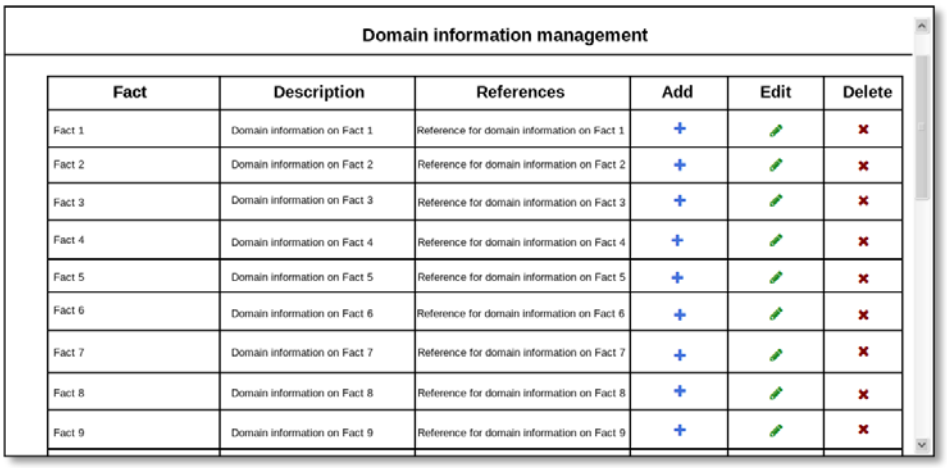


Figure 4.14: Fact list page design

2. *Information on the selected document* can be selected by the user to zoom into. By including information about the selected document, the design adheres to the following design guideline:

    Information about the data itself should be provided.

    The document information is entered into the system by means of a manual process. The manual process requires that the user must gather information about the document, which includes the following:

    - name and surname of the author;
    - name and surname of the author's spouse;
    - birth date of the document author;
    - death date of the document author;
    - document author's location at the start of the Anglo-Boer War;
    - date that the document was written by the author.

    The input page presented to the user is depicted in Figure 4.15. In addition to the information provided by the user through a manual process, the file size of the textual document is populated during the information extraction process discussed in Chapter 3.

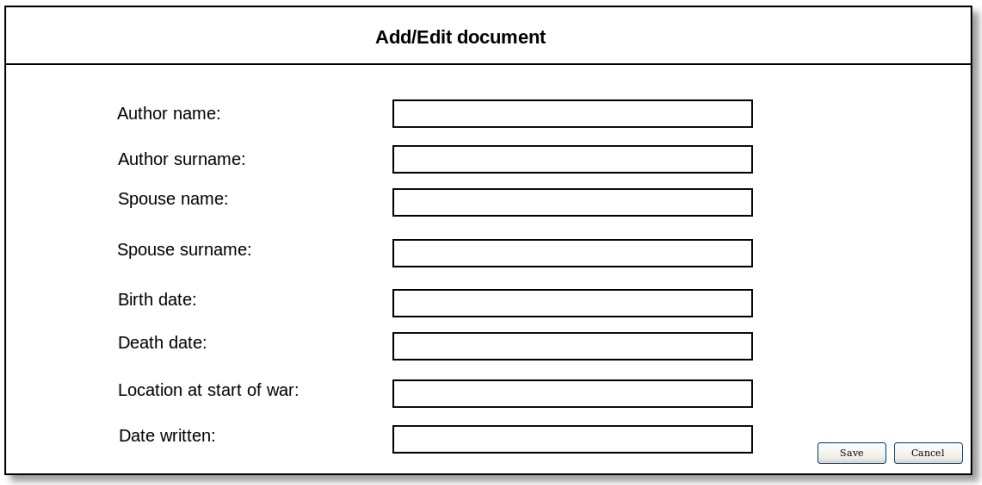


Figure 4.15: Document information Add/Edit page design

All the documents contained in the set are listed and displayed to the user as depicted in Figure 4.16 on the following page. The user can enter document-specific

information by clicking on the *Edit* icon and then subsequently entering the information in the form provided in Figure 4.15.

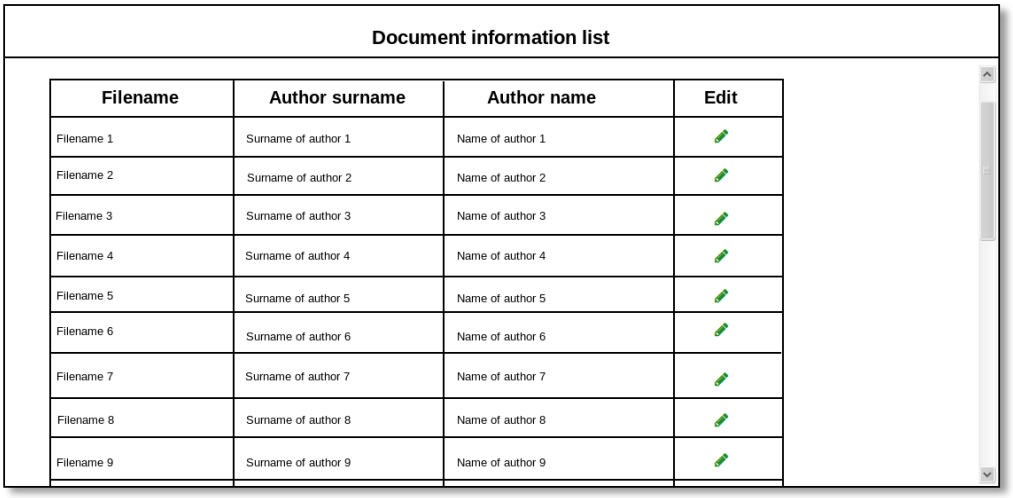


Figure 4.16: Document list page design

3. *Information describing the document set as a whole.* By including information about the source of the document set, the design adheres to the following design guideline:

   Information about the data itself should be provided.

4. The information is populated as part of the database setup and, for the source set used in the study, consists of the following text:

   The Huyser collection contains Anglo-Boer War testimonies of survivors of the named war. The set consists of 41 testimonies written mostly in 1929/1930. Of these, three are not used in this visualisation since they do not describe the theme of the Anglo-Boer War. These testimonies were collected by the publisher, HW Huyser, and were later used as basis for the novel Bittereinders by JRL van Bruggen. The collection has been the focus of recent historical studies and its value has been open for scrutiny by researchers.

Having discussed the design of the visual representation of the coherence of facts in a record set, as well as supporting views to provide additional information to the visualisation, the next section discusses how the design principles and guidelines defined in Section 4.1.3 and listed in Table 4.3 are applied to the design of the visualisation.

4.2.1.1.5 Applying information visualisation principles and guidelines to the design of the visualisation view

The application of each of the design principles and guidelines to the visualisation view is discussed in this section. By ensuring that the design of the interface is based on the principles and guidelines listed in Table 4.3, rigour is applied to the design. Ku *et al*. (2012) advise that principles and guidelines should guide the design of the view and should not be considered as strict rules. Each of the design principles and guidelines, as set out in Section 4.1.3 and listed in Table 4.3, was considered as follows in the design of the visualisation view:

1. Consider providing focus and context user interactions

Focus and context user interactions allow the user to view the detail of information, while being presented with the overall context in which the information exists *(Ku et al*., 2012). Focus and context user interactions are provided in the design of the visualisation by providing the user with the functionality to zoom into a fact as well as the ability to view the context in which the fact appears in a selected document.

The user can select a fact, while at the same time view the document (the context) in which the fact appears.

2. Consider providing details-on-demand user interactions

The user can select to view details-on-demand by either zooming into a document or into a fact. When the user zooms into a fact, details describing the fact are displayed in the *Domain information on selected item* section. In addition, the user can select details on the context in which the fact appears in the document by clicking on the *View in context* button.

When zooming into a document, information about the document is displayed to the user. The user can, in addition, acquire the text of the document on demand by selecting the *Show document* button.

3. Consider providing brushing-and-linking user interactions

Brushing-and-linking user actions are user interactions which cause a change in one visualisation view to change the view of another (Ku *et al*., 2012). Brushing-and-linking interactions are provided in the visualisation by means of changes in a section of the visualisation causing the view of other sections to change. The provided brushing-and-linking interactions are the following:

- Selecting a fact in the *Browsing and filtering* section changes the view in the *Corpus section*, indicating all occurrences of the selected fact in the documents.

- Zooming into a fact by clicking on a fact-representing circle in the *Corpus section*, changes the content of the *Domain information on selected item* to provide domain information on the selected fact. The *Information on selected document section* is simultaneously updated to display information on the document in which the selected fact appears.

- Zooming into a document by clicking on a document-representing circle in the *Corpus section*, changes the content *Information on selected document* section to provide information on the selected document.

4a. Consider providing an overview

An overall view of the whole document set is provided in the *Corpus section*, presenting all the documents as well as all the fact occurrences in the set.

4b. Consider providing zoom functionality

Zoom functionality provided to the user of the visualisation consists of the following:

- Zoom into and out of a document, i.e. the *zoom-into-document view* page.

- Zoom into a fact occurrence in a document, i.e. the *zoom-into-fact* view page.

- Zoom out of a document or fact occurrence to the record set as a whole, i.e. the *overall view* page.

4c. Consider providing functionality to filter out items

Functionality to filter out items is provided by means of the following:

- Dropdown lists in the *Browsing and filtering* section. By selecting a fact in the dropdown list, the user's attention is drawn to all occurrences of the fact in the document set, thereby filtering out all other fact occurrences. The filtering is achieved by changing the colour of all occurrences (indicated by circles inside the document-representing circle) of the selected fact to the colour as indicated next to the dropdown list.

- Zooming-in functionality (either zooming into a document or fact) allows for all other items (documents or facts) to be filtered out.

4d. Consider providing functionality to get details-on-demand

Functionality to provide details-on-demand user interactions was already discussed in Guideline 2. It appears as sub-items under Guideline 4 as well, since it is part of the Shneiderman (1996) user interactions.

4e. Consider providing functionality to show how items relate

The relation between documents and fact occurrences is visually represented by means of circles. Each document is represented by a circle. In each circle, fact occurrences in the document are represented by smaller circles.

4f. Consider providing functionality to keep history of user interactions

For the purpose of the research study, the design does not cater for storing the history of user interactions. Although providing the ability to the user to list and execute previous browsing and filtering actions on facts is useful, it is not considered as cardinal for visually representing the coherence of facts in a record set. It is envisaged that future iterations of the visualisation might include the functionality to store and re-execute user interactions.

4g. Consider providing functionality to extract sub-collections

The *overall view* presents the user with a view of all the occurrences of facts contained in the document set. By zooming into a document, the user is presented with a sub-collection of fact occurrences. The user is also provided with functionality to extract sub-collections by selecting facts in the *Browsing and filtering* section. Sub-collections are visually extracted by changing the colour of the fact occurrences in accordance with the colour indicated by the dropdown list.

5. Consider what the user can see, needs to see and would like to see

What the user can see and needs to see is guided by the objective of the visualisation, namely to visually represent the coherence of facts in a record set.

What the user can see is furthermore restricted by technical shortcomings related to the visualisation medium, which include screen resolution.

What the user would like to see was not known at the start of the design of the visualisation. It was envisaged that the demonstration and evaluation steps in the design science research process (discussed in Chapter 5) could inform the research on what the user would like to see. Future iterations of the design could possibly include elements of what the user would like to see.

6. Additional domain information required by the user should be provided

The *Domain information on selected* item section provides the user with additional domain information. It is suggested by Chen (2005) and Chen *et al*. (2009) that additional domain information assist the user to interpret the visualisation.

The domain information is gathered by the user or domain expert by conducting research on the facts in the document set. The result of the research is then entered into the system by a manual process.

7. Additional information on how to operate the visualisation device should be provided

The design of the visualisation does not cater for providing additional information on how to operate the visualisation. It is assumed that the visualisation is simple and clear enough to operate without requiring instructions. It was envisaged that the assumption could be confirmed or rejected during the demonstration step in the design science research process.

Future iterations of the design could possibility include instructions on how to operate the visualisation in the case where the user indicated user instructions as a requirement.

8. Information on the data itself should be provided

Information on the data is provided to the user as follows:

- Information on the data set as a whole is provided in the *Document set overview* section.
- Each document in the set is considered by the research to be a source of data. Information on each document is presented in the *Information on selected document* section. In addition, the user can select to view the data by clicking on the *Show document* button.

9. The type of information to be visualised should be considered

The design of the visualisation was guided by the requirement that the coherence of domain facts contained in a record set should be visualised.

10. The purpose of the visualisation should be considered

The acquiring of insight into the probable truth value of domain facts contained in a record set was taken into account in the design of the visualisation. The coherence of facts in the document set could be an indication of the possible truth value of the record set.

11. The user profile should be considered

Different views are not being provided to different profiles of users. However, the design took into account that domain experts and other users interested in the domain facts are potential users of the system.

12. The context in which the visualisation will take place should be considered

The design of the visualisation took into account that the visualisation would be presented to the user by means of a software application. Restrictions concerning

limitations with regard to software applications and hardware such as screen resolution and user interactions with the system were considered.

13. The visualisation format should be considered

The visualisation was designed from the start to be a highly interactive system.

14. The mapping of data elements to visual objects should be enhanced by realistic techniques or additional symbols

Data elements are represented by zoomable circles. Fact occurrences are displayed inside the circles, each representing a document. One large circle represents the document set as a whole. The possible use of circular structures to represent data elements was identified by the researcher in the literature study on visualisation techniques discussed in Section 4.1.4.1.

During the process of searching for possible visualisation software tools in the literature as discussed in Section 4.1.4.2, D3.js was identified as a possible software tool to implement a software solution. In the investigation of D3.js as possible software tool, the researcher identified Mike Bostock's (Bostock, 2013b) *Zoomable circle-packing* visualisation technique (depicted in Figure 4.17) as a possible starting point for designing a solution to the research problem.

It was envisaged that the demonstration and evaluation steps in the design science research process could inform the research on whether the chosen visualisation technique is realistic and whether the technique solves the problem statement.

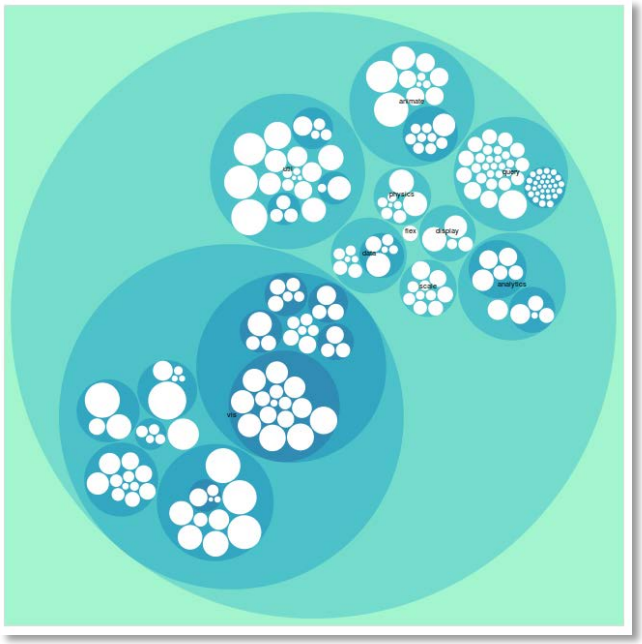


Figure 4.17: Zoomable circle-packing technique (Bostock, 2013b)

15. Consider the use of minimal user actions to accomplish a task

The use of minimal user actions was considered in the design of the visualisation in order to keep the design simple.

16. Consider providing a number of ways to achieve the same goal

The design of the visualisation does not provide a number of ways to achieve the same goal. It is proposed by the research that providing more than one way to achieve a goal would add unnecessary complexity to the visualisation, which would not result in a clean and simple visualisation as set out in the following guideline:

> *Visualisations should be clean and simple.*

17. Consider functionality to provide redo/undo of actions

It is proposed by the research that the interactions provided in the visualisation are simple enough to ignore providing redo/undo of actions.

18. The organisation of the visual layout should be considered

The layout of the visualisation was carefully considered. The following sections of information were provided in the layout of the interface:

- *Corpus* section;
- *Browsing and filtering* section;
- *Information on selected document* section;
- *Domain information on selected item* section;
- *Document set overview* section.

19. Design choices should be consistently applied to the solution

Care was taken to apply design choices throughout the solution.

20. The cognitive load on the user should be minimal

In order to ensure that the cognitive load on the user was kept to a minimum, the following design choices were made:

- Only allow the user to select up to four facts in the *Browsing and filtering* section. Making a decision about the coherence of facts in the record set becomes more difficult when more than four facts are selected.

- Do not display all the author names on each document-representing circle, because it overloads the visualisation. Instead, the tooltip element is to be displayed with the name of the author whenever the cursor is moved over the document-representing circle.

- Display only what the user needs to see. The display of any unnecessary information might lead to a cognitive overload.

21. Provide the user with information on alternatives when several actions are available

Alternative actions to achieve the same goal are not provided by the design, because it will add unnecessary complexity to the visualisation.

22. Consider removing distracting information from the visualisation solution

Care is taken by the design of the visualisation to only provide information to the user that is required to make an assumption on the coherence of the facts contained in the record set.

23. Consider means to reduce the data set

Means to reduce the data set was not considered in the design of the visualisation because the visual representation is appropriate for the data set provided by the textual data source, i.e. the Huyser collection of testimonies (Huyser, n.d.). Future application of the solution to a larger data source might require the design to consider means to reduce the data set.

24. Maintain a visual entity budget

For the purpose of visualising the Huyser collection of testimonies (Huyser, n.d.), the need to only display a certain number of documents or facts is not applicable. Future application of the solution to a larger data source might require the design to consider only displaying a certain number of documents and facts. Alternatively, the design can be altered to display only documents containing the facts selected in the *Browsing and filtering* section.

25. Visual aggregates should visually summarise the underlying data

Visual aggregates are presented to the user on the following two levels:

- All the documents contained in the document set are visually aggregated inside one circle.

- Each document is a visual of the underlying fact occurrences in the document.

26. Visualisations should be clean and simple

Care was taken in the design of the visualisation to keep the design clean and simple, in particular when following the guidelines below:

- *Consider what the user can see, needs to see and would like to see.*
- *Consider the use of minimal user actions to accomplish a task.*
- *Consider providing a number of ways to achieve the same goal.*
- *Consider functionality to provide redo/undo of actions.*
- *Provide the user with information on alternatives when several actions are available.*
- *Consider removing distracting information from the visualisation solution.*

27. Visual aggregates should be discriminatingly distinguishable from data items

The visual representation is designed in such a manner that the document set representing circles, the document-representing circles and the fact occurrence-representing circle are all distinguishable by being represented by different colours and sizes.

28. Keep in mind that visual abstractions might be misleading

Although misleading visual abstractions are possible, the user must be assisted in acquiring insight into the selected fact in the *Browsing and filtering* section by providing him or her with the following information:

- researched domain information in the *Domain information on selected item* section;
- information describing the document (author details, date written and size details) in which the selected fact occurs;
- visual information on the coherence of the selected fact in the document set.

29. Visual aggregates should be interpretable

A visualisation must be interpretable to create an insight. The main insight that the visualisation is designed to achieve is insight into the coherence of facts contained in the record set.

The demonstration and evaluation steps in the design science research process could inform the research on whether the user considers the visual aggregates to be interpretable.

30. Shortcomings related to the visualisation tools should be considered

Shortcomings related to the technologies and tools used to implement the visualisation were considered in the development (discussed in Section 4.2.2) of the visualisation application.

31. Shortcomings related to the mapping of concepts with representations should be considered

Mapping documents and facts to circular representations was considered as an appropriate technique to visually represent the concepts. It was assumed that shortcomings related to the mapping of concepts with the representations would be identified and handled during the development (discussed in Section 4.2.2) of the visualisation application.

32. Shortcomings related to technical problems to represent information should be considered

Shortcomings to screen size and resolution were considered in the design of the visualisation. For optimal viewing, a screen size of 23 inches and a screen resolution of 1920 x 1080 pixels should be used.

Having discussed the view to visually represent the coherence of facts in a record set, the next section briefly discusses the design of the database schema to support the view.

4.2.1.2 Design of the database schema

The database schema design (depicted in Figure 4.18 on the following page) to support the design of the user view discussed in Section 4.2.1.1 is presented in this section.

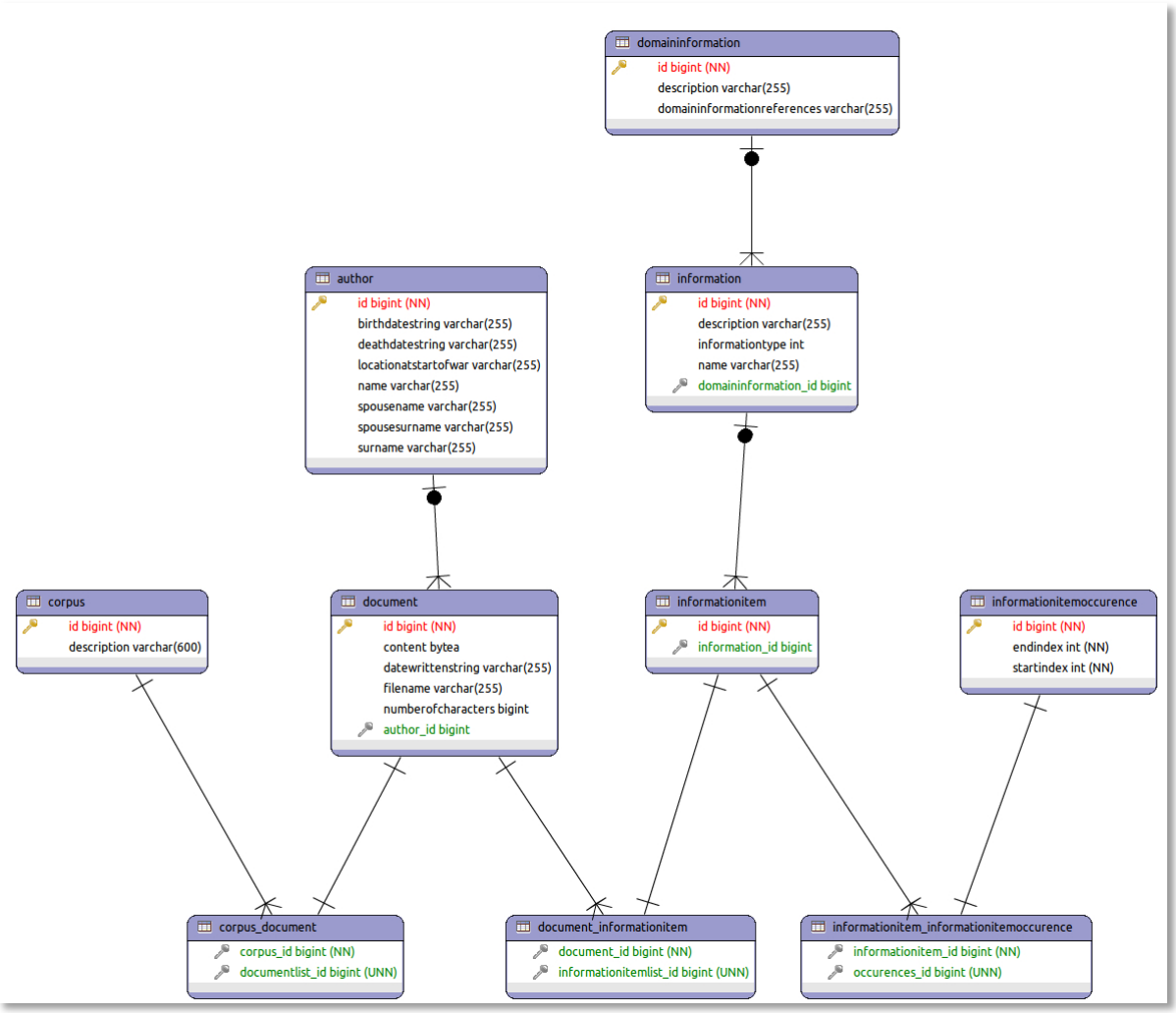


Figure 4.18: Database schema design

The schema defines the database entities that will be used to store the information extracted from the textual document set (discussed in Chapter 3), as well as document and domain information entered into the system by the user (discussed in Section 4.2.1.1.4) and basic information (information that describes the source of the documents).

The database schema design consists of the following entities:

- *corpus:* contains a description of the document source; populated during the information extraction process described in Chapter 3.
- *document:* contains information on each document in the set, including the filename, the size of the file in number of characters, the date the original document was written and a reference to the author of the document. The information (filename, file content, size of the file and the author reference) is populated during the information extraction

process described in Chapter 3, as well as being manually entered by the user (discussed in Section 4.2.1.1.4).

- *author:* referenced from the *document* entity, containing information about each of the document authors. The entity is populated by the user manually entering the information (discussed in Section 4.2.1.1.4).

- *corpus_document:* an entity that links each document to the document set.

- *informationitem:* each fact appearing inside a document is represented by the *Informationitem* entity. Each fact can have multiple occurrences inside the document; each occurrence is represented by the *informationitemoccurrence* entity.

- *information:* contains information about the fact represented by the *informationitem* entity, which *includes* a description and a name of the fact. An additional information type variable is defined for future use to enable the system to define whether the fact being described is an entity or an event.

- *domaininformation:* contains information gathered by the user through domain research and entered into the system by means of a manual process (discussed in Section 4.2.1.1.4).

- *document_informationitem:* an entity that links each fact represented in the visualisation to the document it is contained in.

- *informationitemoccurrence:* each fact contained in a document can appear multiple times in the document. The fact is only shown once in the document-representing circle. Each occurrence of the fact inside the document is stored in the *informationitemoccurrence* entity, enabling the system to highlight the location of each occurrence to the user.

- *informationitem_informationitemoccurrence:* an entity that links each occurrence of a fact to its representation in a document-representing circle.

Having discussed the design of both the user view to visually represent the coherence of facts contained in a record set and the database schema to support the view, the next section gives an overview of the activity of building the designed visualisation interface.

4.2.2 Building a visualisation interface from the design

The activity of building the visualisation interface designed in Section 4.2.1 is discussed in this section. The focus of the discussion is on the technologies employed. The actual development process is not discussed because the process is deemed too insignificant to contribute to the research debate. In addition, screenshots of the developed application (with mockup data used) are also provided in this section.

4.2.2.1 Technologies used in the implementation of the design

The software development skill set of the researcher influenced the decision of the researcher to implement the design as an interactive web application according to the architecture as depicted in Figure 4.19.

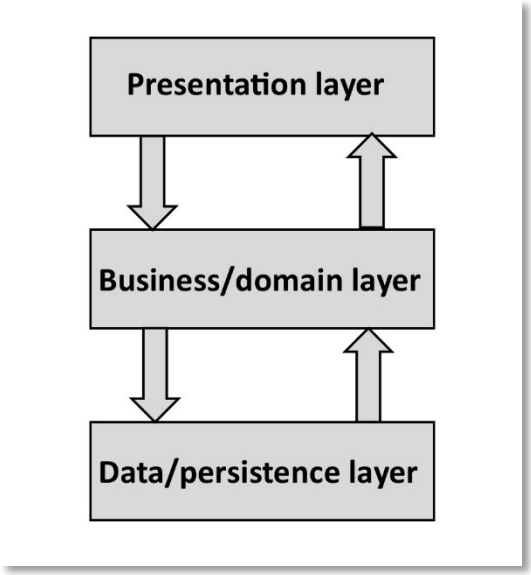


Figure 4.19: Three-tier application architecture

An interactive application can be built on a three-tier architecture (Ameller, Collell, & Franch, 2013; Fowler, 2002; Grozev & Buyya, 2014; Marston, 2012) as depicted in Figure 4.19. Each of these tiers, also known as layers, is responsible for performing a different role as listed in Table 4.6.

Table 4.6: Roles of layers in the three-tier architecture

Layer	Role/responsibility
Presentation	Represents the user interface; handles the user view and user input.
Business/domain	Handles the main application logic.
Data/persistence	Handles access to the database.

Each of the layers in the architecture communicates to either a higher-level or a lower-level layer by means of requests and responses.

The *Presentation layer,* for instance, displays a prompt for input from the user. The input received from the user is sent by means of a request to the *Business layer*. The *Business layer* logic code determines what to do with the input and returns a response that is presented to the user by means of the *Presentation layer*.

If need be, the *Business layer* can send a request for data to the *Data layer*. The *Data layer* next retrieves the data from the database and sends the data via a response to the *Business layer*. Figure 4.20 depicts the technologies used in each of the architecture layers in order to implement and instantiate the interactive web application. The decision to use the specific technologies is based on the researcher's own practical skillset and is therefore considered as the quickest and most effective manner to implement and instantiate the solution.

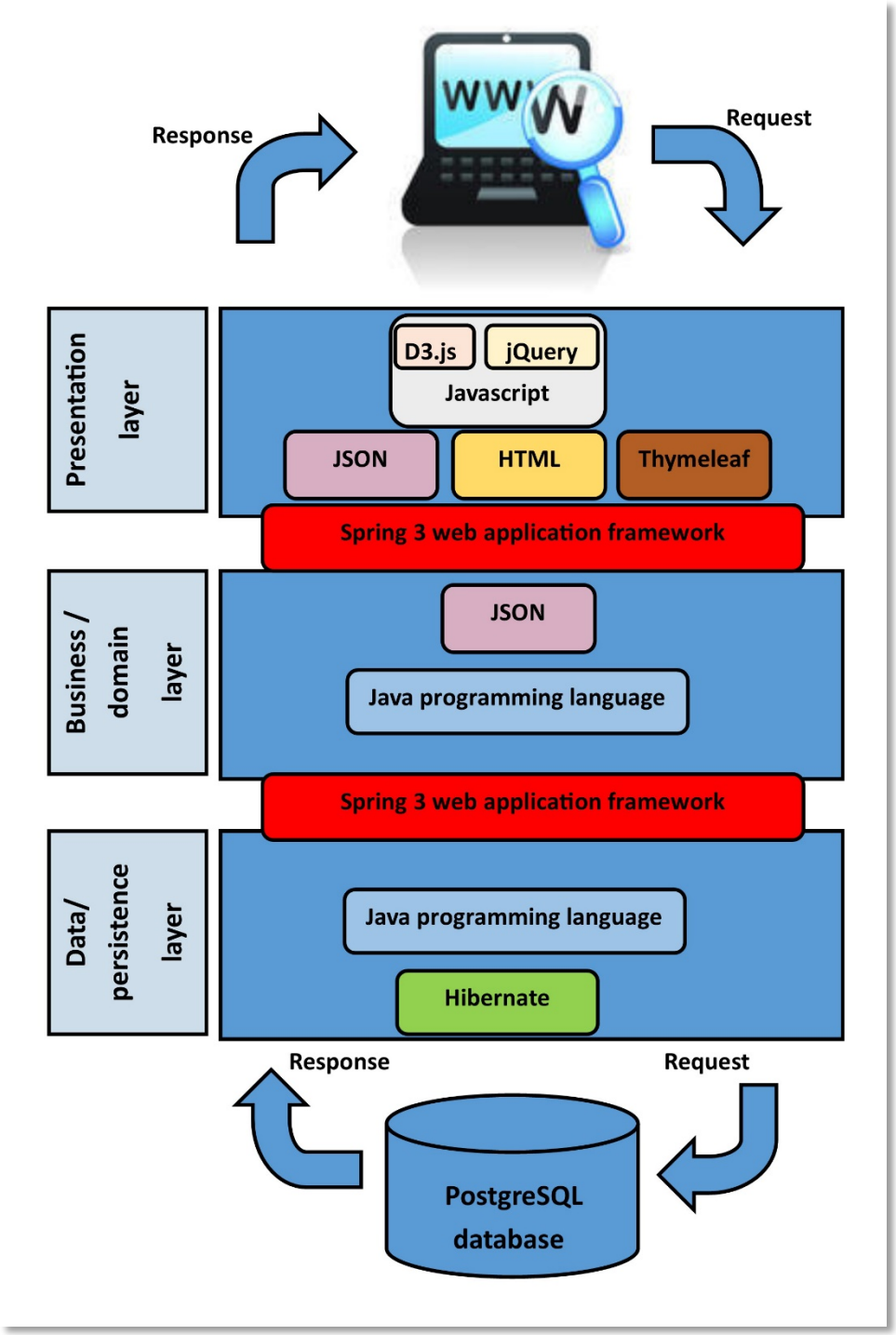


Figure 4.20: Technologies used in each layer

Each of the technologies used in the architecture layers is discussed as follows:

- Javascript, jQuery, HTML and Thymeleaf are technologies used to develop the user interface, allowing the user to interact with the system through the use of a web browser. These technologies are used for presenting the user with a view as well as the ability to perform any user actions.

- D3.js is a visualisation framework developed in the Javascript language. The framework allows for a number of visualisation techniques, including the zoomable circle-packing example provided by Bostock (2013b). The technique provided by Bostock (2013b) was used as starting point for the development of the visualisation solution. Additional functionality on top of this technique was added to allow for the implementation of the design discussed in Section 4.2.1.

- The data (documents and fact information) used as input to the visualisation is provided to the user interface in the XML data representation format. XML as technology appears in both the *Presentation* and the *Business layer* since the XML data is constructed in the *Business layer* and utilised in the *Presentation layer*.

- The Java programming language is used to programmatically implement the decision-making rules and responses to user input in the *Business layer.* The construction of the XML data representation format is also implemented in the Java programming language.

- Hibernate is a framework written in the Java programming language, assisting with database query and retrieval requests.

- According to Vosloo and Kourie (2008), a web application framework is "a collection of software components that support the development and execution of Web-based user interfaces" (p 1). Spring 3 is such a web application framework providing reusable software components that are used on all three of the architecture layers.

- The application is hosted on a Tomcat web application server.

Having provided an overview of the technologies used to implement the design of the visualsation, screenshots (using mockup data) of some of the implemented pages are presented next.

4.2.2.2 Screenshots (with mockup data) of the developed interface

Screenshots of the developed interface pages are presented in this section. In order to keep the screens simple, the pages are presented with mockup data. The mockup data was also used in order to keep the development of the information extraction artifact and the visual representation artifact separate, thereby simplifying the development of each artifact.

In order to provide a starting page from where the user can access the visualisation interface, the document management views and the domain information management views, the page depicted in Figure 4.21 was developed. This page displays a simple word cloud constructed from all the words contained in all of the textual documents. It is proposed that the word cloud is useful in getting an idea of the domain described by the facts contained in the documents. In addition, access is provided by this page to the domain information management and document information management as the main visualisation pages.

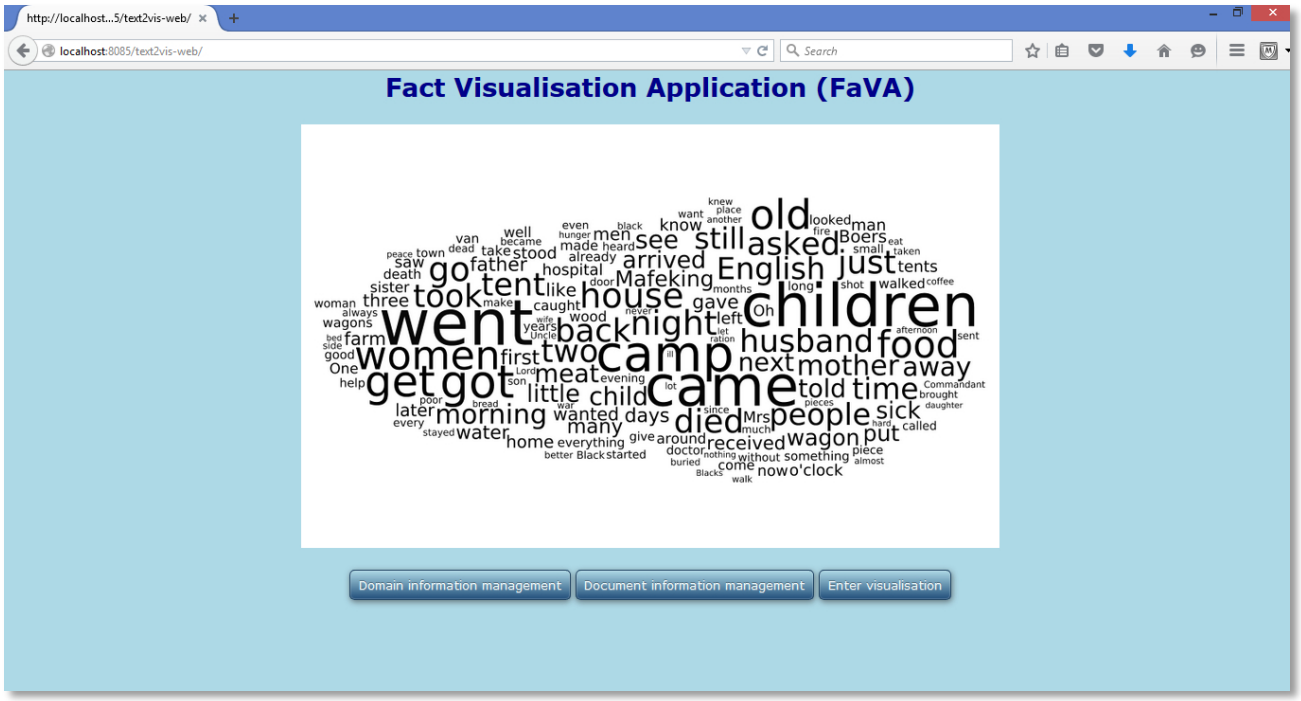


Figure 4.21: Starting page

The main visualisation interface is depicted in Figure 4.22 on the following page. The mockup data contains eight documents and 53 information occurrences in the documents.

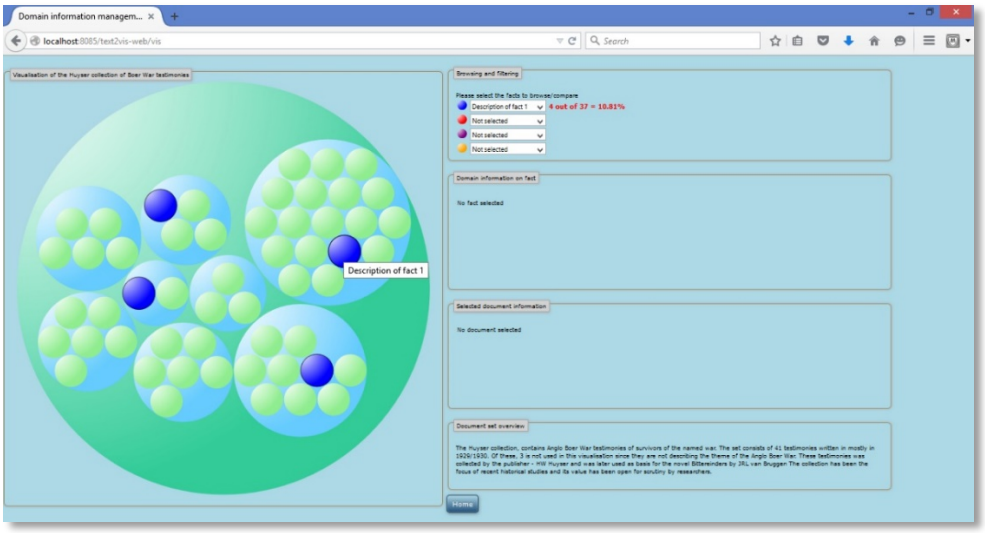


Figure 4.22: Main visualisation interface

Zooming into a fact, results in the view depicted in Figure 4.23 are presented to the user.

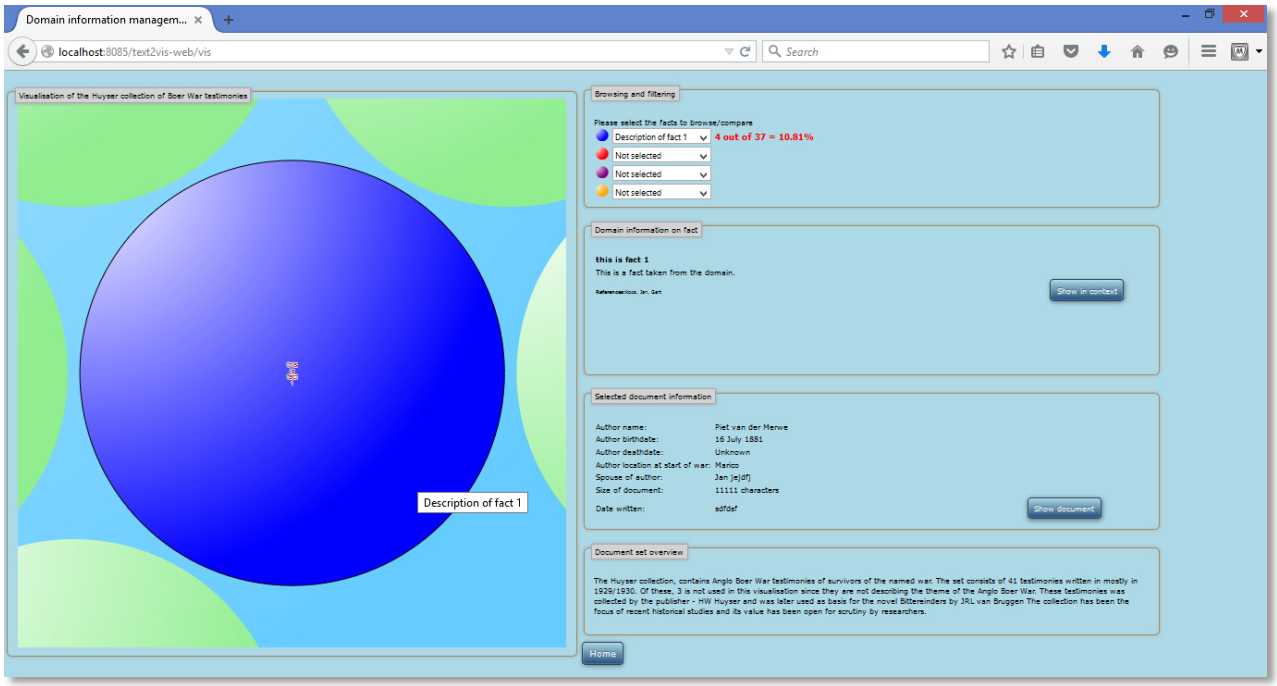


Figure 4.23: Zoom-into-fact view

Upon clicking on the *Show in context* button (after zooming into a fact), all the occurrences of the fact in the document are highlighted and presented to the user as depicted in Figure 4.24.

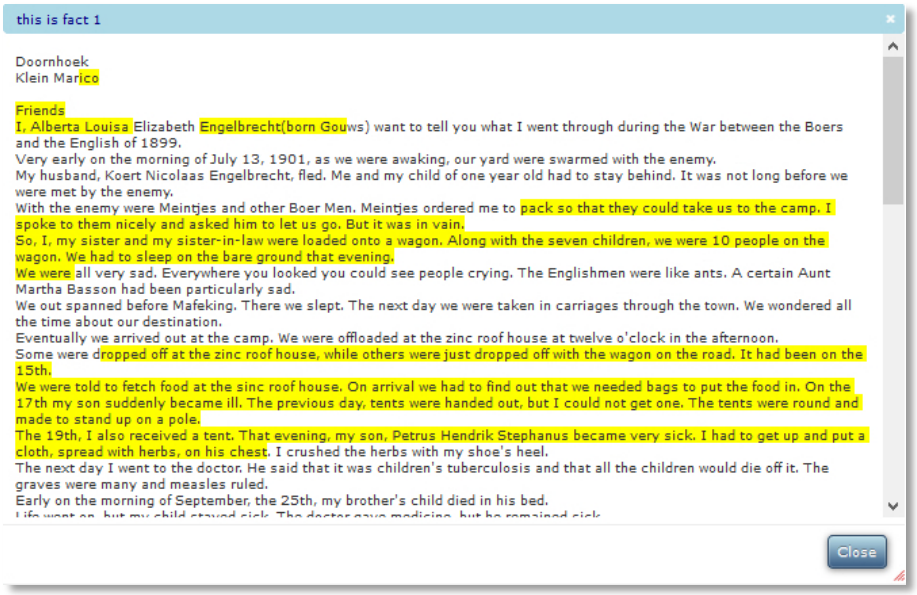


Figure 4.24: Show in context dialog

Choosing to zoom into a document results in the user being presented with the view depicted in Figure 4.25.

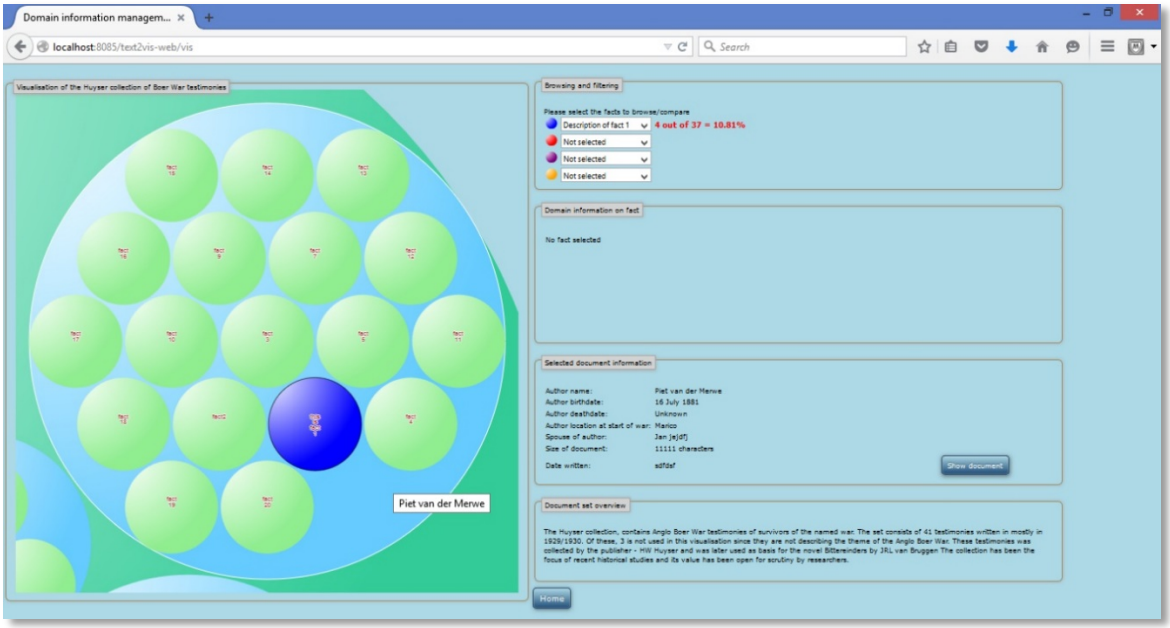


Figure 4.25: Zoom-into-document view

Upon choosing from the starting page to manage domain, the user is presented with the page depicted in Figure 4.26.

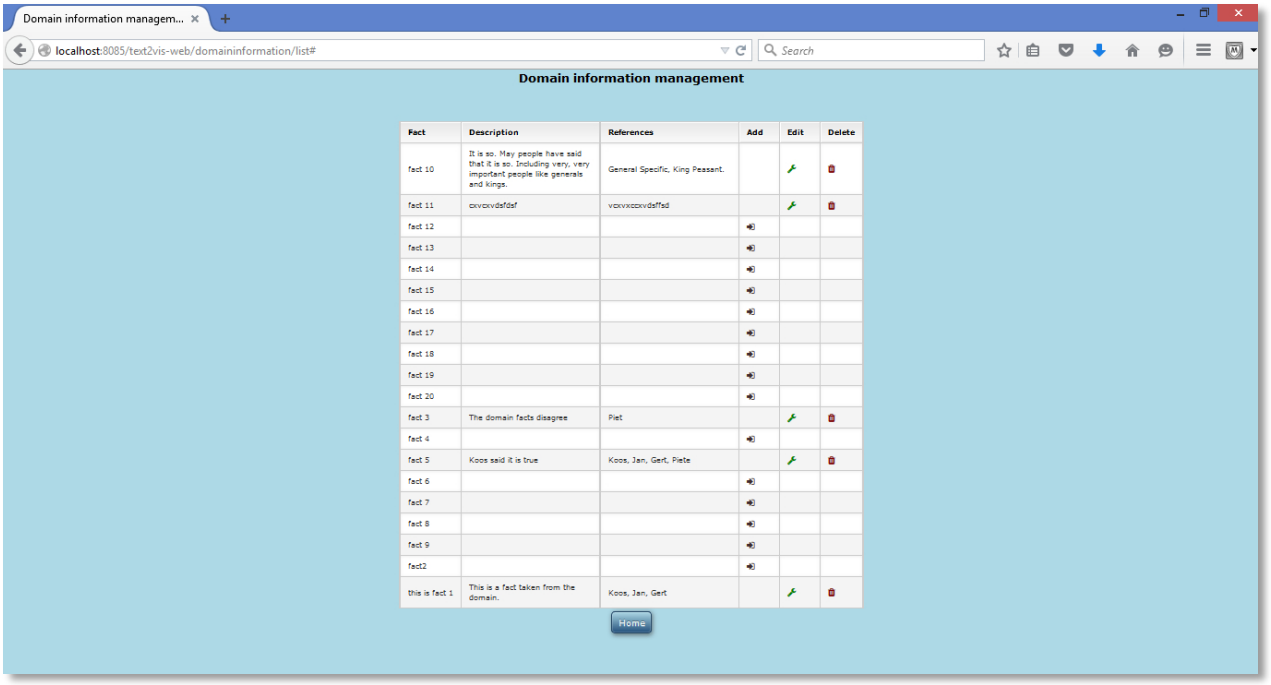


Figure 4.26: Domain information management page

The dialog view presented to the user when he or she chooses to add or edit domain information is depicted in Figure 4.27.

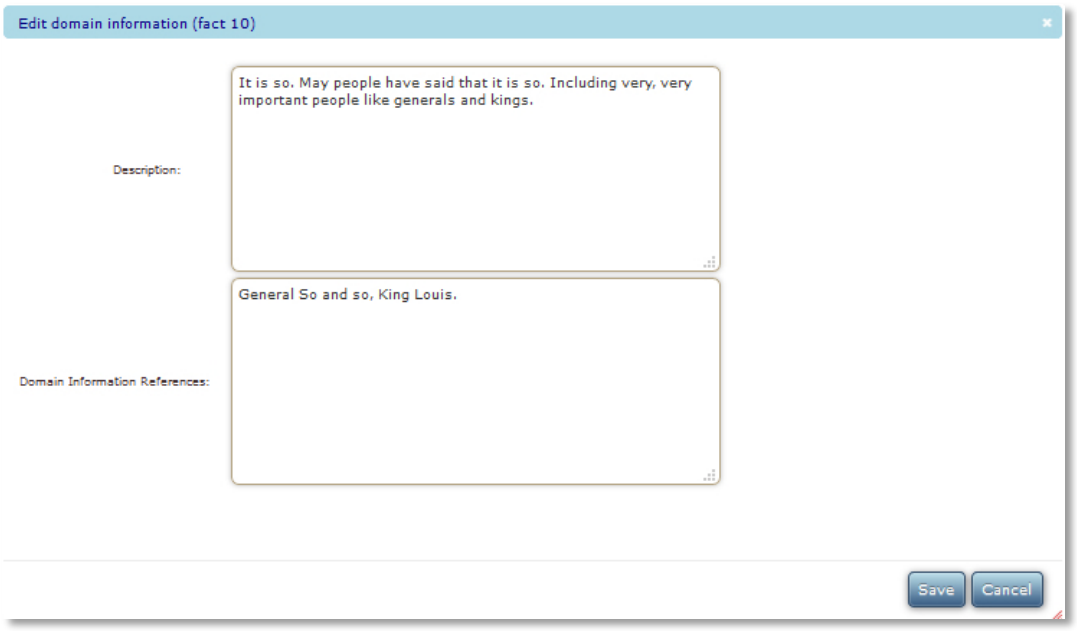


Figure 4.27: Add/Edit domain information dialog view

Upon choosing from the starting page to manage domain information, the user is presented with the page depicted in Figure 4.28.

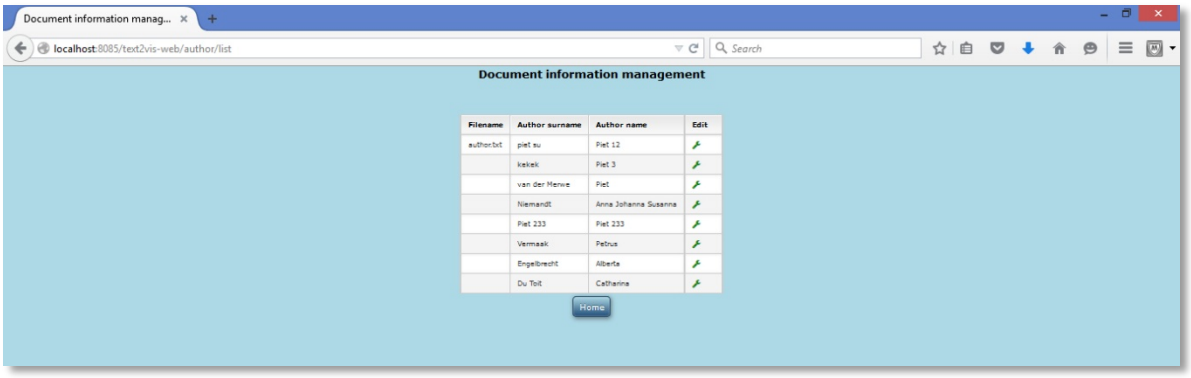


Figure 4.28: Document information management page

The dialog view presented to the user when he or she chooses to add or edit document information is depicted in Figure 4.29.

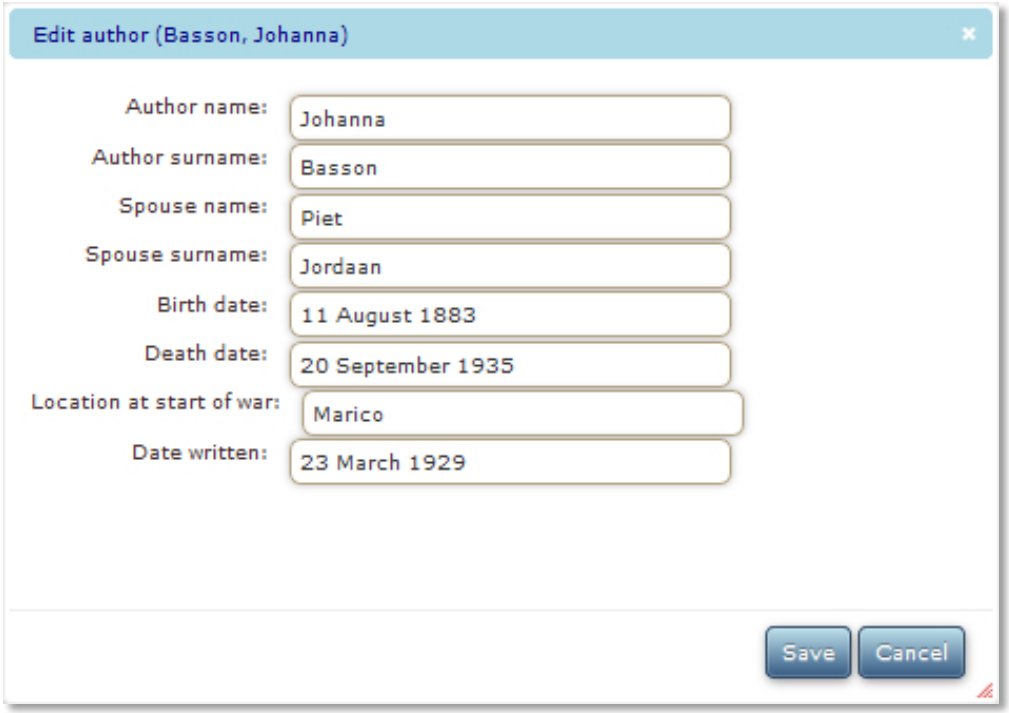


Figure 4.29: Add/Edit document dialog view

A screenshot of the overall view after populating the database with the extracted information is depicted in Figure 4.30 on the following page. In the screenshot, the user selected to view

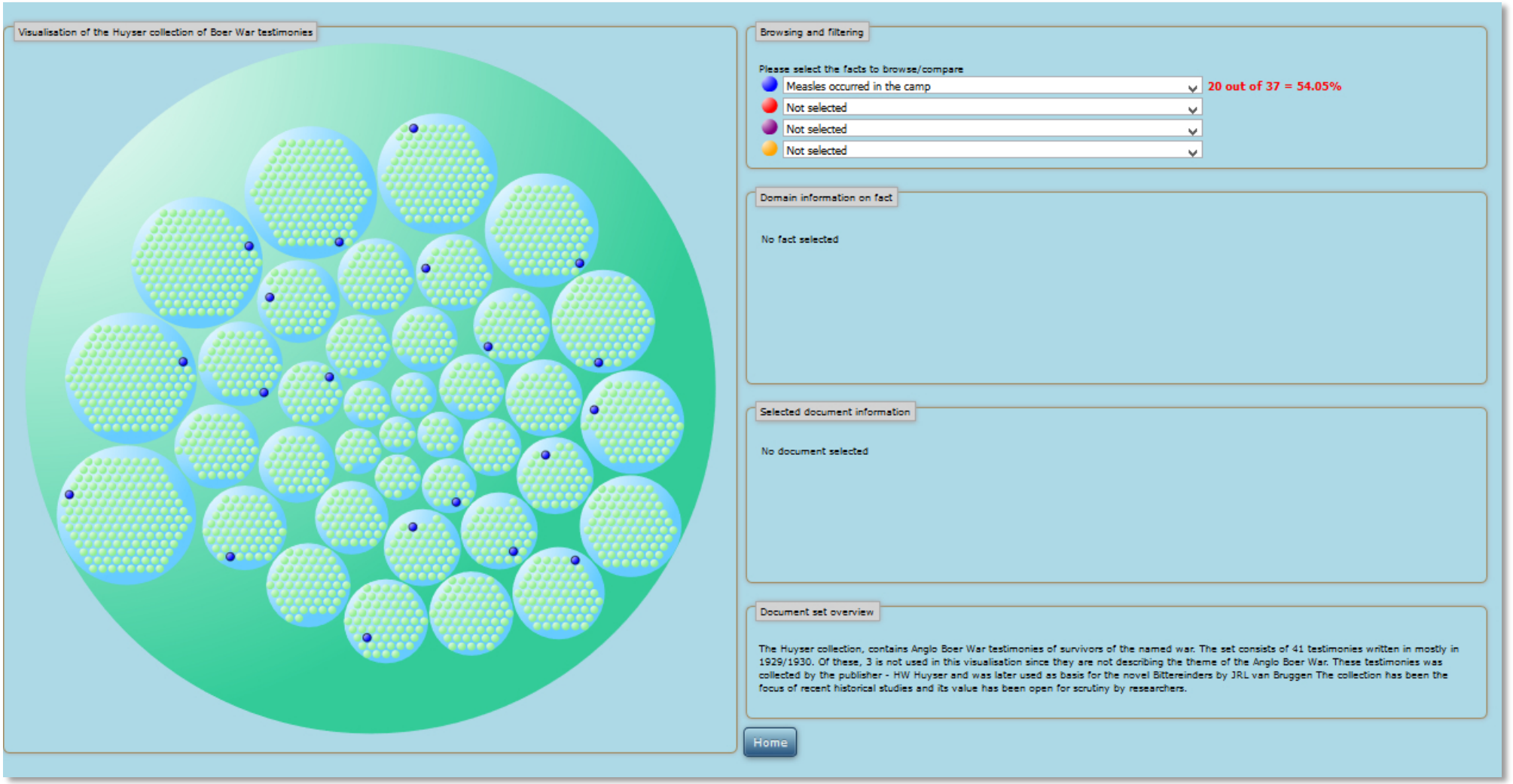


Figure 4.30: Coherence of the "measles occurred in camp" fact in the document set

the coherence of the “measles occurred in the camp” fact in the document set. Each occurrence of the fact in a document is visually represented by colouring its fact-representing circles with the colour as indicated next to the dropdown list (dark blue). Additional information is also given to the user – the fact is described in 20 out of 37 documents (54.05%). Therefore, the user might make the assumption that there is coherence of the “measles occurred in the camp” fact occurrences in the document set.

In this section, a proposed software solution to visually represent the coherence of facts in a document set was provided. The design and building activities required to construct and instantiate the solution were discussed in the previous sections. It is postulated that the development of an *intermediate* model to visually represent the coherence of facts in a document set can be informed by the specific software solution. Such a model is proposed in the next section.

4.3 SUMMARY

An *intermediate* model (depicted in Figure 4.31) for visual representation of the coherence of facts is presented next by reflecting on the visual representation components identified in the previous section.

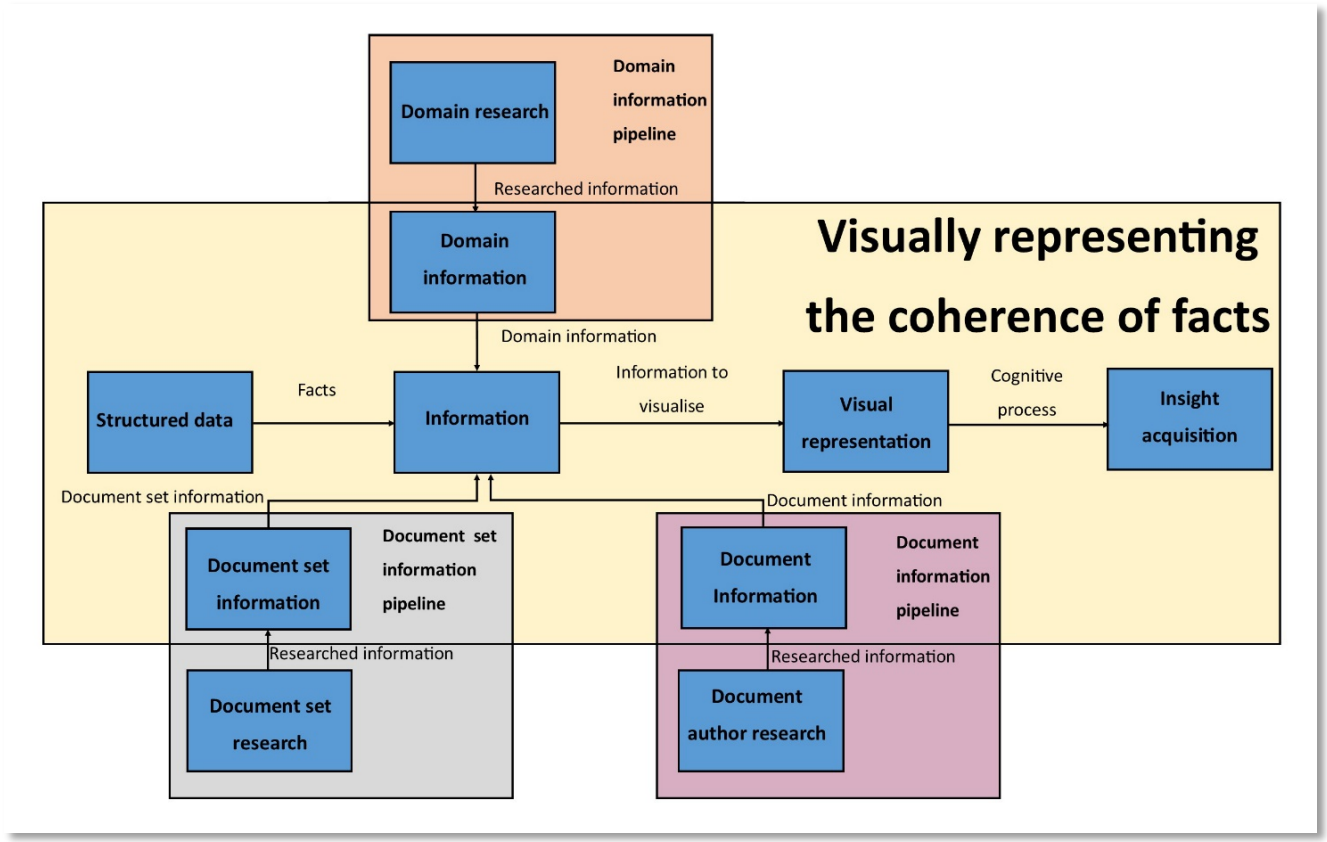


Figure 4.31: Intermediate model - visually representing the coherence of facts in a structured data form

From the preceding discussion of the proposed design of the visualisation view and the instantiation of the visualisation artifact, it is proposed that the following components can be combined into a model (depicted in Figure 4.31 on the previous page) to visually represent the coherence of facts contained in a document set:

- *Information* to be visually represented. The information is provided by the structured data containing domain facts, the domain information pipeline, the document source information pipeline and the document set information pipeline.

- *Structured data* containing domain fact information. The structured data is produced by an information extracting process discussed in Chapter 3.

- *Domain information pipeline* consisting of the following components:

    - *Domain research* (process external to the visualisation artifact) on domain facts being visualised.

    - *Domain information* management including the addition, deletion and update of domain information describing a fact.

- *Document source information pipeline* consisting of the following components:

    - *Document author research* (process external to the visualisation artifact) on the authors of the document containing the domain facts.

    - *Document information* management including the addition, deletion and update of document information describing a document containing domain facts.

- *Document set information pipeline* consisting of the following components.

    - *Document set research* (process external to the visualisation artifact) with regard to information about the source of the document set.

    - *Document set information* management. The document set information is added to the visualisation during the visualisation project setup as described in Section 3.2.3.

- *Visual representation* of the coherence of facts in a record set.

- *Insight acquisition* by means of cognitively processing the visual representation.

Having discussed a proposed *intermediate* software solution and model for visually representing the coherence of facts contained in a document set, the next chapter provides a discussion of the development and demonstration of the *final* software solution as well as the validation of the *final* model.

CHAPTER 5: DEVELOPMENT, DEMONSTRATION AND EVALUATION OF THE FINAL SOFTWARE SOLUTION AND VALIDATION OF THE FINAL MODEL

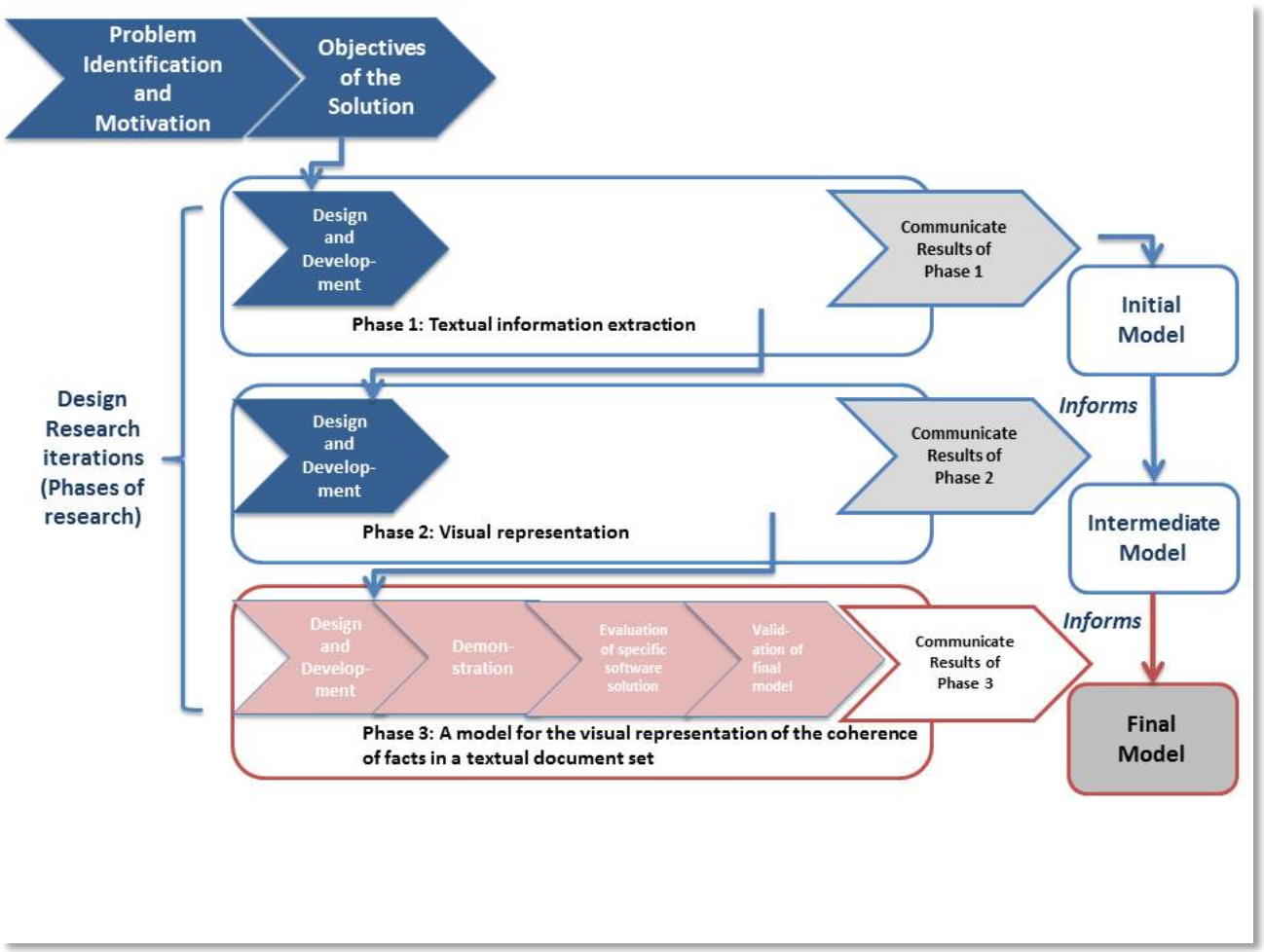


Figure 5.1: Current phase in research – final Information Visualisation model

As can be seen in Figure 5.1, Chapter 5 describes the design and development of the *final* model for the visual representation of the coherence of facts in a textual document set. The chapter also documents the demonstration and evaluation of the specific *final* software solution and the validation of the final model.

The specific *final* software solution was developed by combining instantiations of the *initial* information extraction (Phase 1) and the *intermediate* visual representation solutions (Phase 2). The results of Phase 1 and Phase 2 of the research process were models (*initial* and *intermediate*) whose development were informed by the instantiation of the respective specific *initial* and *intermediate* solutions. The development of *final* model for the visual representation of the coherence of facts in a textual document set, was in turn informed by

the combined *initial* information extraction and *intermediate* visual representation solution instantiations and models.

In order to ensure rigour (Peffers *et al*., 2007), the *final* specific solution and *final* model was demonstrated and evaluated. The demonstration and evaluation of the *final* specific software solution and final model is are discussed in the next sections.

5.1 FINAL SOFTWARE SOLUTION

In this section the demonstration (Section 5.1.1) and evaluation (Section 5.1.2) of the *final* software solution to visually represent the coherence of facts in a textual document are discussed.

5.1.1 Final software solution demonstration

According to Peffers *et al*. (2007), the demonstration activity of the design science research model involves the demonstration of the built solution's ability to solve the identified problem by any suitable means. For the purpose of the research study, the instantiation of the *final* software solution (including the *initial* information extraction and the *intermediate* visual representation parts) was demonstrated by means of individual and group (no more than five individuals at a session) demonstration sessions. As a result the *final* solution was demonstrated to five experts in the domain (Anglo-Boer War history) of the information being visualised as well as five IT experts in order to get feedback from both domain and technical users.

The decision to use five experts per field of expertise is based on Nielsen and Landauer's (1993) proposition that between three and five expert users are adequate for conducting user evaluations. The participant profiles are summarised in Table 5.1. In order to adhere to ethical considerations, the identity of each participant is protected.

Table 5.1: Final solution demonstration participant profiles

Description of participant	Level of expertise	Participant category
Doctorate student in the field of Anglo-Boer War history	Doctorate student level	History expert
Publisher of historical books, including Anglo-Boer War material	Excellent	History expert
Professional Anglo-Boer War researcher	Senior	History expert
Professional Anglo-Boer War researcher	Senior	History expert
Professional Anglo Boer-War researcher	Senior	History expert
Experienced software developer	Senior	IT expert
IT expert	Senior	IT expert
IT researcher	Intermediate	IT expert

Description of participant	Level of expertise	Participant category
IT software engineer	Senior	IT expert
IT business analyst	Master's level student	IT expert

The demonstrations were done during four interview sessions: two individual interview sessions with history experts, one group interview session with three history experts and one interview session with five IT experts. Each interview session lasted for about two hours.

At the start of each demonstration session, each participant was presented with a consent form, which presented the participant with a description of the research topic, the research process and potential benefits of the research. The participant was furthermore presented with a confidentially and withdrawal clause indicating that the participant's identity was protected and was treated as confidential. The participant gave his or her consent to participate in the research upon signing of the consent form.

Each demonstration session consisted of the following:

- an introduction to the research topic;
- a presentation of the visual representation artifact. By using different walkthrough scenarios, the functionality and the use of the interface were explained in detail;
- a presentation of how the information was extracted from the document source, which included a description of the document preparation and fact-extracting processes. The use of ATLAS.ti to extract facts from textual documents was included in the presentation;
- each participant was given time to use the visual interface to explore the facts in the document set and to experience the process of acquiring some insight into the facts contained in the documents.

After each demonstration session, the participants were presented with a demonstration feedback questionnaire containing the information and questions listed in Table 5.2 on the following page.

Table 5.2: Final solution demonstration feedback questionnaire

Background

My name is Louis Engelbrecht, and as part of my Master of Computer Science degree at the University of South Africa, a software solution artifact was developed.

The aim of the software artifact is to visually represent the domain facts contained in a set of textual documents. For the purpose of this research, the Huyser collection of Anglo-Boer War testimonies is used as source for the visualisation.

After being provided with a demonstration of the software, you will be requested to evaluate the visualisation by completing this form. Your input will be used to improve/enhance the software artifact.

Participant information

1. What is your involvement with the history of the Anglo-Boer War or in IT?

2. What is your level of expertise on the history of the Anglo-Boer War or in IT?

Instructions

1. Please complete the next section by marking your choice with an "X" in the box provided.
2. Please answer all questions.
3. Please note that each answer only requires a single choice.

Evaluation form

Please read the statements with regard to the demonstrated software below and how strongly you agree or disagree. Please use the following rating system:

Strongly disagree	Somewhat disagree	Neither agree or disagree	Agree	Strongly agree
1	2	3	4	5

Number	Statement	Rating				
1.	The technique for visualising the facts contained in the testimonies is appropriate.	1	2	3	4	5
2.	The visualisation technique employs a minimal set of actions to accomplish a task.	1	2	3	4	5
3.	Functionality to control the level of details is available.	1	2	3	4	5
4.	The visual layout of the visualisation is effective.	1	2	3	4	5
5.	The cognitive load on the user is minimal.	1	2	3	4	5
6.	The visualisation contains extracted information.	1	2	3	4	5
7.	There is too much information on the visualisation view.	1	2	3	4	5
8.	Additional functionality to show less testimonies and facts should be considered.	1	2	3	4	5
9.	The visualisation is clean and simple.	1	2	3	4	5
10.	Testimonies/documents and facts are easily distinguishable from each other.	1	2	3	4	5
11.	The visualisation leads to misleading interpretations.	1	2	3	4	5
12.	The visualisation is interpretable.	1	2	3	4	5
13.	The coherence of facts contained in the Huyser collection is represented in the visualisation.	1	2	3	4	5
14.	The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history.	1	2	3	4	5

Comments/Recommendations:

_____

_____

_____

_____

_____

_____

The statements presented to the participants were carefully defined in order to be used as input for the evaluation of the *final* instantiation of the solution as discussed in Section 5.1.2. Each statement is used to contribute towards evaluating the following:

1. whether the instantiation of the solution contributes to:

    - identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form;
    - identifying components of a process to visually represent the coherence of facts in a textual;

2. whether design principles and guidelines (provided in Section 4.1.3) were considered in the design of the *intermediate* visual representation solution. It is proposed that these statements can also be used to enhance/improve the *final* solution;

3. whether the instantiation of the *final* solution has utility, i.e. if it is useful. The comments and recommendations sections, in addition, contribute to evaluating the solution's usefulness.

For each statement, the participant had to indicate how strongly he or she agreed or disagreed by selecting a value from one to five according to a five-point Likert scale (Likert, 1932). Table 5.3 contains a list of the feedback questionnaire statements, each with an indication of its contribution to the artifact evaluation.

Table 5.3: Contributions of feedback statements to the evaluation of the final solution

Statement	Evaluation contribution
1. The technique for visualising the facts contained in the testimonies is appropriate.	Design principles and guidelines were considered.
2. The visualisation technique employs a minimal set of actions to accomplish a task.	Design principles and guidelines were considered.
3. Functionality to control the level of details is available.	Design principles and guidelines were considered.
4. The visual layout of the visualisation is effective.	Design principles and guidelines were considered.
5. The cognitive load on the user is minimal.	Design principles and guidelines were considered.

Statement	Evaluation contribution
6. The visualisation contains extracted information.	Contributes to identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form
7. There is too much information on the visualisation view.	Design principles and guidelines were considered.
8. Additional functionality to show less testimonies and facts should be considered.	Design principles and guidelines were considered.
9. The visualisation is clean and simple.	Design principles and guidelines were considered.
10. Testimonies/documents and facts are easily distinguishable from each other.	Design principles and guidelines were considered.
11. The visualisation leads to misleading interpretations.	Design principles and guidelines were considered.
12. The visualisation is interpretable.	Design principles and guidelines were considered.
13. The coherence of facts contained in the Huyser collection is represented in the visualisation.	Contributes to identifying components of a process to visually represent the coherence of facts in a textual
14. The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history.	The instantiation of the solution has utility.

In addition to the statements presented in the evaluation form, a section for comments and recommendations was provided. The comments and recommendations received from the participants can be used for future re-iteration to the design and build step in the design science research model in order to enhance and improve the instantiations of the solution. For the purpose of the research due to time constraints and the positive feedback from the demonstration audience, the solution is deemed as suitable without the need to re-iterate the design and build steps.

The results from the feedback questionnaire were used as input into the artifact evaluation step in the design science research model as discussed in the next section.

5.1.2 Final software solution evaluation

A discussion of the evaluation of the instantiation of the *final* solution produced by the research study is provided in this section. The evaluation is guided by the responses received from the experts (Anglo-Boer War and IT experts) after the demonstration of the solution. The demonstration consisted of a walkthrough of descriptive scenarios to indicate the use of the *final* software solution. The analysis of the data was done afterwards using *Microsoft Excel*. The feedback was collected in the form of a five-point Likert scale questionnaire.

To analyse the data, the researcher used pivot tables and charts for each of the statements presented in the feedback questionnaire. The responses from the IT experts were furthermore compared with the responses from the Anglo-Boer War experts. The statements presented in the feedback questionnaire were designed in order to determine the following:

- Whether the instantiation of the solution contributes to:
    - identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form;

- identifying components of a process to visually represent the coherence of facts in a document set;

- Were design principles and guidelines considered in the design of the artifact in order to ensure rigour?

- Whether the solution is useful?

The next section analyses the feedback data in order to determine whether the solution contributed to achieving the objectives of the research.

5.1.2.1 Feedback statements with regard whether the solution contributes towards identifying process components

Feedback statements with regard to determining whether the solution contributed identifying process components are analysed in this section as follows.

1. The visualisation contains extracted information.

To evaluate whether the solution contributes to identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form, responses to the following statement were considered:

- *The visualisation contains extracted information.*

As indicated in the pivot table (Table 5.4) and illustrated in the bar graph on the following page (Figure 5.2), all the Anglo-Boer War experts strongly agreed with the statement. Two IT experts agreed and three IT experts strongly agreed with the statement.

Table 5.4: The visualisation contains extracted information

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	0	5
IT	0	0	0	2	3
Grand Total	**0**	**0**	**0**	**2**	**8**

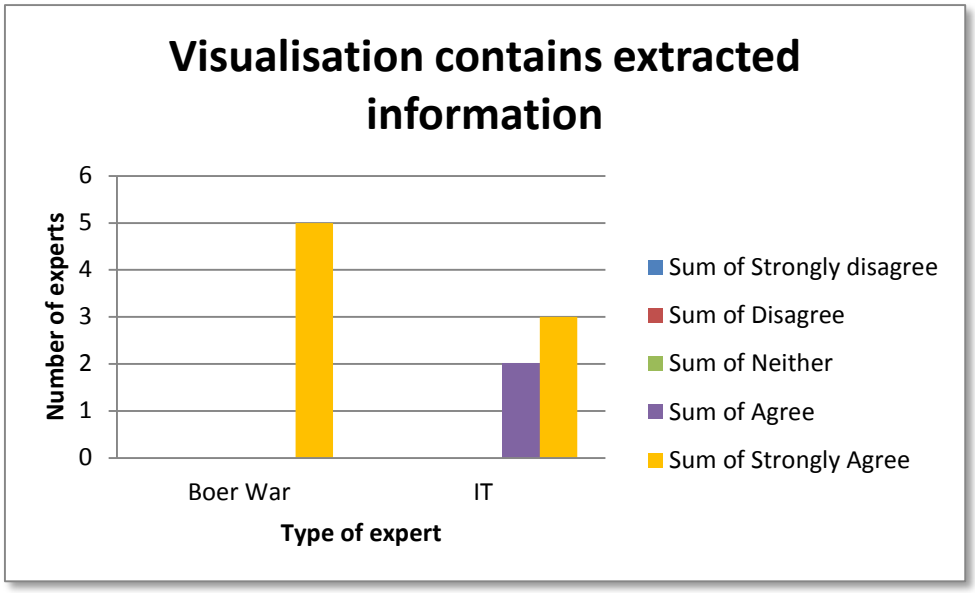


Figure 5.2: Visualisation contains extracted information

In general, 80% of the experts strongly agreed with the statement, while 20% agreed. Thus, it could be concluded that the visualisation contained extracted information. Therefore, the solution contributes towards identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form.

2. The coherence of facts contained in the Huyser collection is represented in the visualisation.

To evaluate whether the solution contributes to identifying components of a process to visually represent the coherence of facts in a document set, responses to the following statement were considered:

- *The coherence of facts contained in the Huyser collection is represented in the visualisation.*

As indicated in the pivot table (Table 5.5) and illustrated in the bar graph (Figure 5.3) on the following page, all the Anglo-Boer War experts strongly agreed with the statement. One IT expert agreed, three IT experts strongly agreed and one IT expert indicated that he or she neither agreed nor disagreed with the statement.

Table 5.5: The coherence of facts contained in the Huyser collection is represented in the visualisation

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	0	5
IT	0	0	1	1	3
Grand Total	**0**	**0**	**1**	**1**	**8**

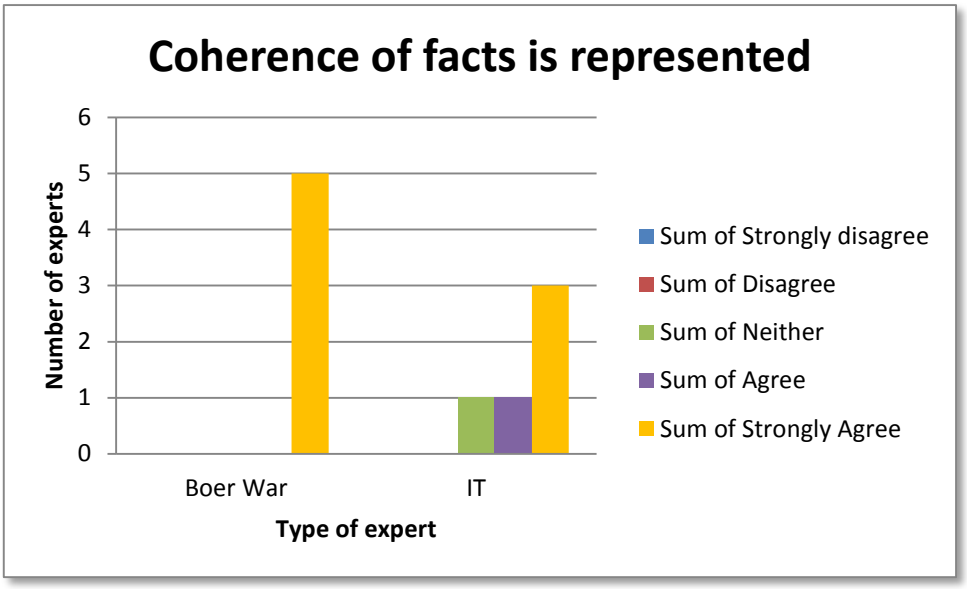


Figure 5.3: Coherence of facts is represented

In general, 80% of the experts strongly agreed with the statement, while 10% agreed. Thus, it could be concluded that the coherence of facts contained in the Huyser collection was represented in the visualisation. Therefore, the solution contributes towards identifying components of a process to visually represent the coherence of facts in a document set.

Furthermore, it is concluded in this section that, from the analysis of the feedback received from the demonstration participants, the instantiation of the solution contributed identifying the information extraction as well as the visual representation process components.

The next section analyses the feedback data in order to determine whether the solution design applied information visualisation design principles and guidelines.

5.1.2.2 Feedback statements with regard to the application of design principles and guidelines

An analysis of the feedback data is provided in this section in order to determine whether the solution design applied information visualisation design principles and guidelines.

1. The technique for visualising the facts contained in the testimonies is appropriate.

In order to evaluate whether the technique for visualising the facts contained in the testimonies was appropriate, responses to the following statement were considered:

- *The technique for visualising the facts contained in the testimonies is appropriate.*

As indicated in the pivot table (Table 5.6) and illustrated in the bar (Figure 5.4), two Anglo-Boer War experts agreed and three Anglo-Boer War experts strongly agreed with the statement. One IT expert agreed and four IT experts strongly agreed with the statement.

Table 5.6: The technique for visualising the facts contained in the testimonies is appropriate

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	2	3
IT	0	0	0	1	4
Grand Total	**0**	**0**	**0**	**3**	**7**

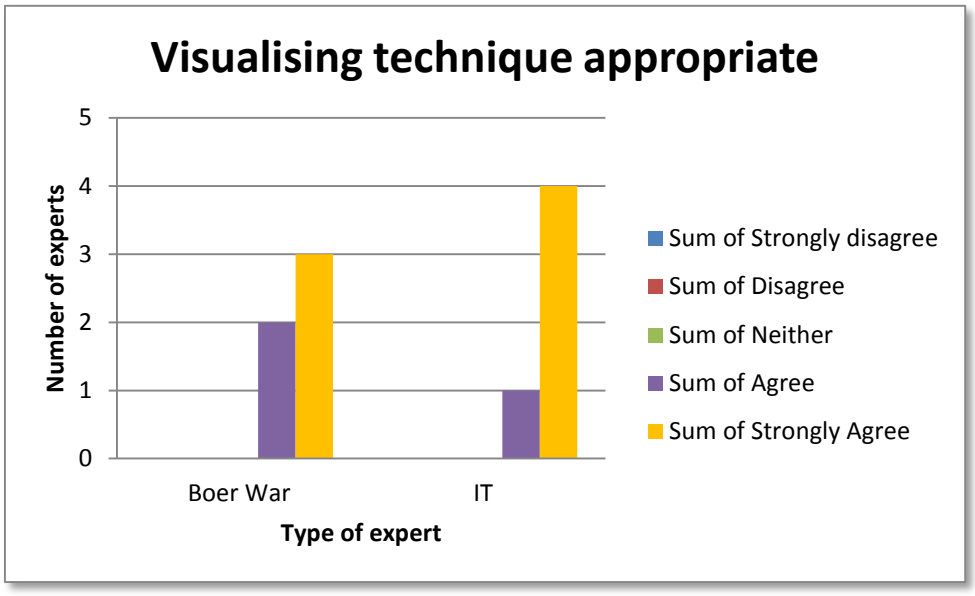


Figure 5.4: Visualising technique appropriate

In general, 70% of the experts strongly agreed with the statement, while 30% agreed. Thus, it could be concluded that the technique for visualising the facts contained in the testimonies was appropriate.

2. The visualisation technique employs a minimal set of actions to accomplish a task.

In order to evaluate whether the visualisation technique employed a minimal set of actions to accomplish a task, responses to the following statement were considered:

- *The visualisation technique employs a minimal set of actions to accomplish a task.*

As indicated in the pivot table (Table 5.7) and illustrated in the bar graph (Figure 5.5) on the following page, two Anglo-Boer War experts agreed and three Anglo-Boer War experts

strongly agreed with the statement. One IT expert indicated that they neither agreed nor disagreed, one IT expert agreed and three IT experts strongly agreed with the statement.

Table 5.7: The visualisation technique employs a minimal set of actions to accomplish a task

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	2	3
IT	0	0	1	1	3
Grand Total	**0**	**0**	**1**	**3**	**6**

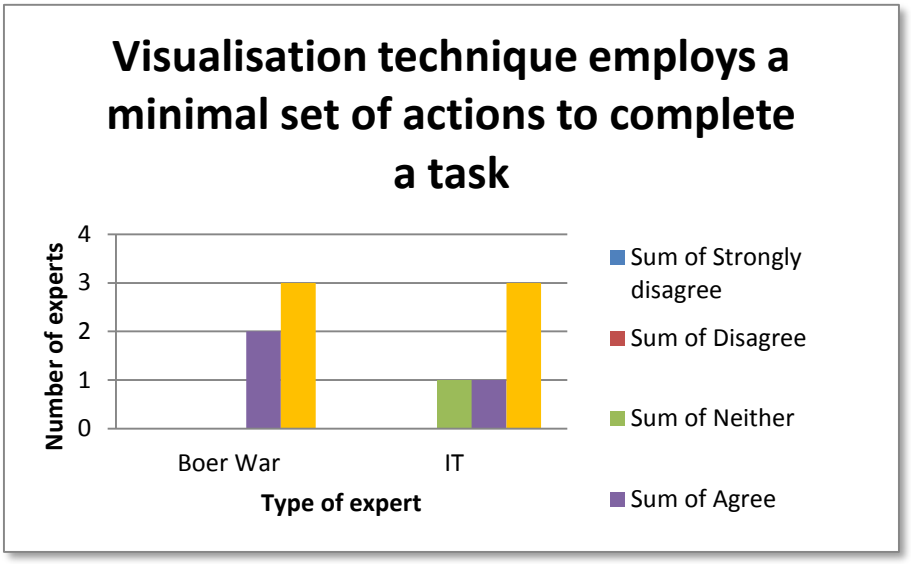


Figure 5.5: Visualisation technique employs a minimal set of actions to complete a task

In general, 60% of the experts strongly agreed with the statement, while 30% agreed. Thus, it could be concluded that the visualisation technique employed a minimal set of actions to complete a task.

3. Functionality to control the level of details is available.

In order to evaluate whether functionality to control the level of details was available, responses to the following statement were considered:

- *Functionality to control the level of details is available.*

As indicated in the pivot table (Table 5.8) and illustrated in the bar graph (Figure 5.6) on the following page, one Anglo-Boer War expert agreed and four Anglo-Boer War experts strongly agreed with the statement. One IT expert indicated that he or she neither agreed nor disagreed, three IT experts agreed and two IT experts strongly agreed with the statement.

Table 5.8: Functionality to control the level of details is available

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	1	4
IT	0	0	1	2	2
Grand Total	**0**	**0**	**1**	**3**	**6**

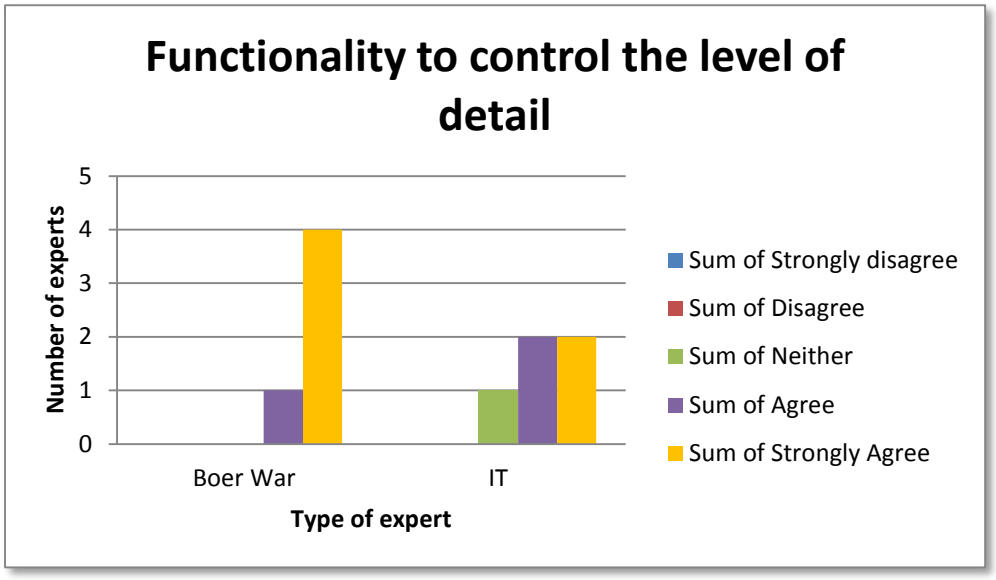


Figure 5.6: Functionality to control the level of detail

In general, 60% of the experts strongly agreed with the statement, while 30% agreed. Thus, it could be concluded that functionality to control the level of details was available.

3. The visual layout of the visualisation is effective to gain an insight into the coherence of facts in the document set.

In order to evaluate whether the visual layout of the visualisation was effective to gain an insight into the coherence of facts in the document set, responses to the following statement were considered:

- *The visual layout of the visualisation is effective.*

As indicated in the pivot table (Table 5.9) and illustrated in the bar graph (Figure 5.7) on the following page, one Anglo-Boer War expert agreed and four Anglo-Boer War experts strongly agreed with the statement. Two IT experts agreed and three IT experts strongly agreed with the statement.

Table 5.9: The visual layout of the visualisaiton is effective

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	1	4
IT	0	0	0	2	3
Grand Total	**0**	**0**	**0**	**3**	**7**

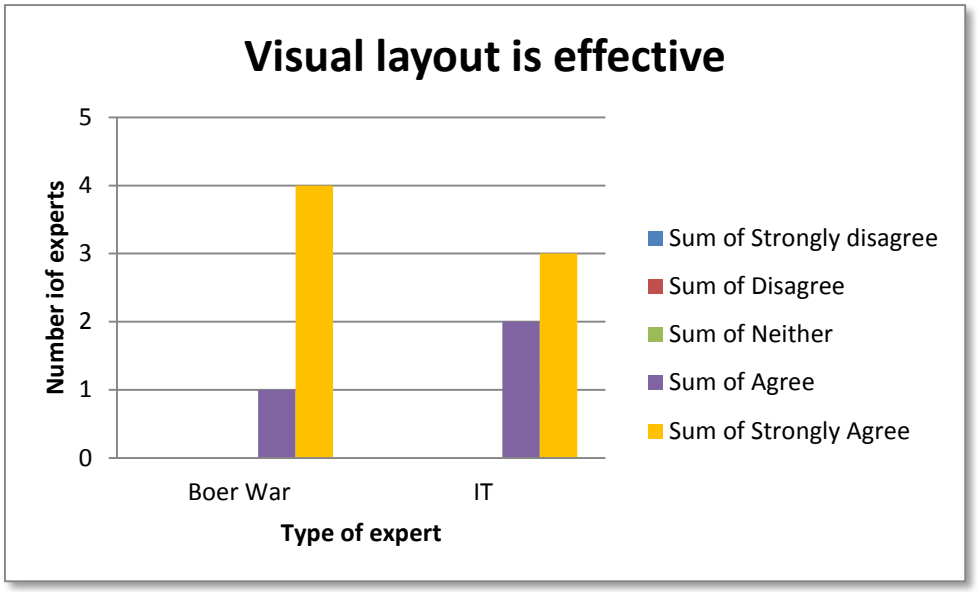


Figure 5.7: Visual layout is effective

In general, 70% of the experts strongly agreed with the statement, while 30% agreed. Thus, it could be concluded that the visual layout of the visualisation was effective.

4. The cognitive load on the user is minimal.

In order to evaluate whether the cognitive load on the user was minimal, responses to the following statement were considered:

- *The cognitive load on the user is minimal.*

As indicated in the pivot table (Table 5.10) and illustrated in the bar graph (Figure 5.8) on the following page, one Anglo-Boer War expert agreed and four Anglo-Boer War experts strongly agreed with the statement. One IT expert indicated that he or she neither agreed nor disagreed with the statement, three IT experts agreed and one IT expert strongly agreed with the statement.

Table 5.10: The cognitive load on the user is minimal

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	1	4
IT	0	0	1	3	1
Grand Total	**0**	**0**	**1**	**4**	**5**

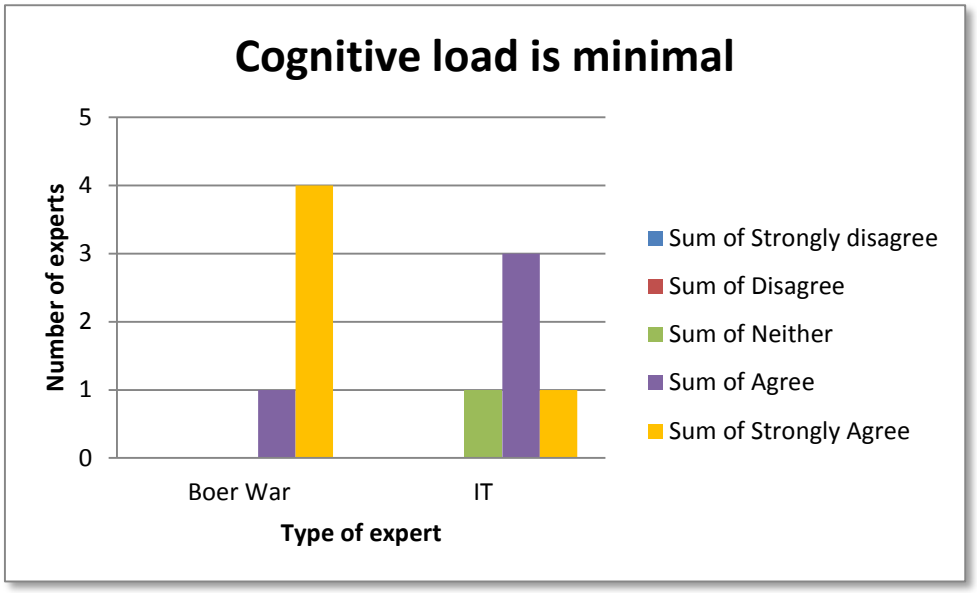


Figure 5.8: Cognitive load is minimal

In general, 50% of the experts strongly agreed with the statement, while 40% agreed. Thus, it could be concluded that the visual layout of the visualisation was effective.

5. There is too much information on the visualisation view.

In order to evaluate whether there was too much information on the visualisation view, responses to the following statement were considered:

- *There is too much information on the visualisation view.*

As indicated in the pivot table (Table 5.11) and illustrated in the bar graph (Figure 5.9) on the following page, three Anglo-Boer War experts strongly disagreed, one Anglo-Boer War expert disagreed and one Anglo-Boer War expert indicated that they neither agreed nor disagreed with the statement. One IT expert strongly disagreed, two IT experts disagreed and two IT experts indicated that they neither agreed nor disagreed with the statement.

Table 5.11: There is too much information on the visualisation view

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	3	1	1	0	0
IT	1	2	2	0	0
Grand Total	**4**	**3**	**3**	**0**	**0**

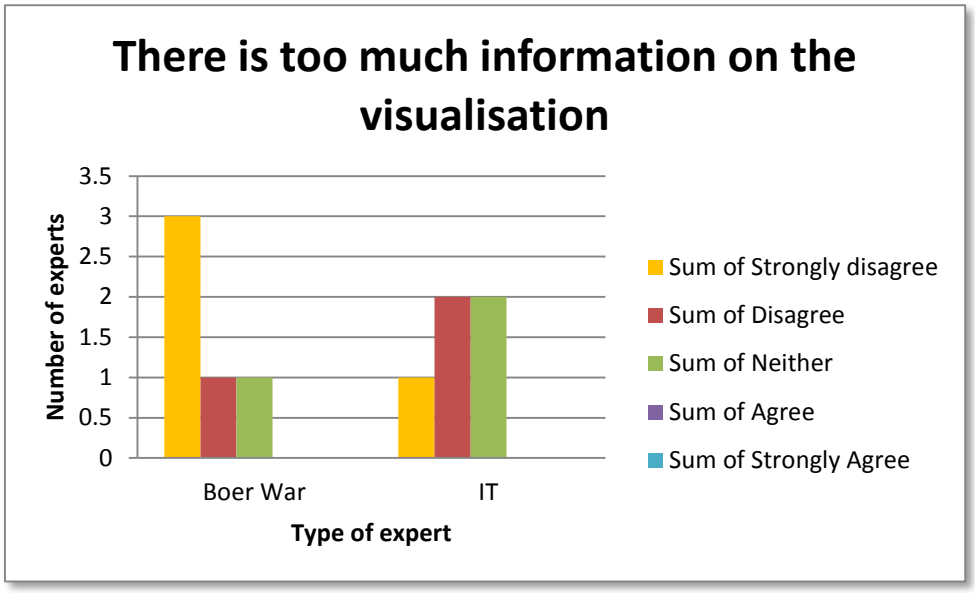


Figure 5.9: There is too much information on the visualisation

In general, 40% of the experts strongly disagreed with the statement, while 30% disagreed, which indicates that a total of 70% of respondents disagreed with the statement. Thus, it could be concluded that there was **NOT** too much information on the visualisation.

6. Additional functionality to show less testimonies and facts should be considered.

In order to evaluate whether additional functionality to show less testimonies and facts should be considered, responses to the following statement were considered:

- *Additional functionality to show less testimonies and facts should be considered.*

As indicated in the pivot table (Table 5.12) and illustrated in the bar graph (Figure 5.10) on the following page, three Anglo-Boer War experts strongly disagreed and one Anglo-Boer War expert indicated that he or she neither agreed nor disagreed with the statement. One IT expert disagreed, two IT experts indicated that they neither agreed nor disagreed with the statement and two IT experts strongly agreed with the statement.

Table 5.12: Additional functionality to show less testimonies and facts should be considered

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	3	0	1	1	0
IT	0	1	2	0	2
Grand Total	**3**	**1**	**3**	**1**	**2**

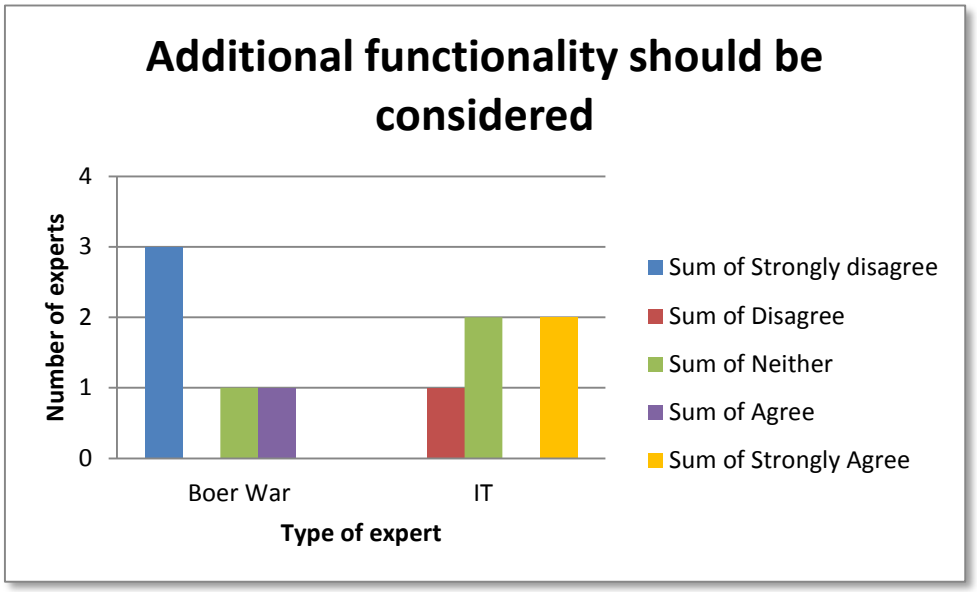


Figure 5.10: Additional functionality should be considered

In general, 30% of the experts strongly disagreed with the statement, while 10% disagreed and 20% strongly agreed. Although the majority of the respondents disagreed with the statement, it is proposed that the item could be reconsidered, given the wide distribution of responses. The researcher cannot conclude with certainty that it is not necessary to add additional functionality to show less testimonies and facts.

7. The visualisation is clean and simple.

In order to evaluate whether the visualisation was clean and simple, responses to the following statement were considered:

- *The visualisation is clean and simple.*

As indicated in the pivot table (Table 5.13) and illustrated in the bar graph (Figure 5.11) on the following page, four Anglo-Boer War experts strongly agreed and one Anglo-Boer War expert agreed with the statement. One IT expert agreed and four IT experts strongly agreed with the statement.

Table 5.13: The visualisation is clean and simple

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	1	4
IT	0	0	0	1	4
Grand Total	**0**	**0**	**0**	**2**	**8**

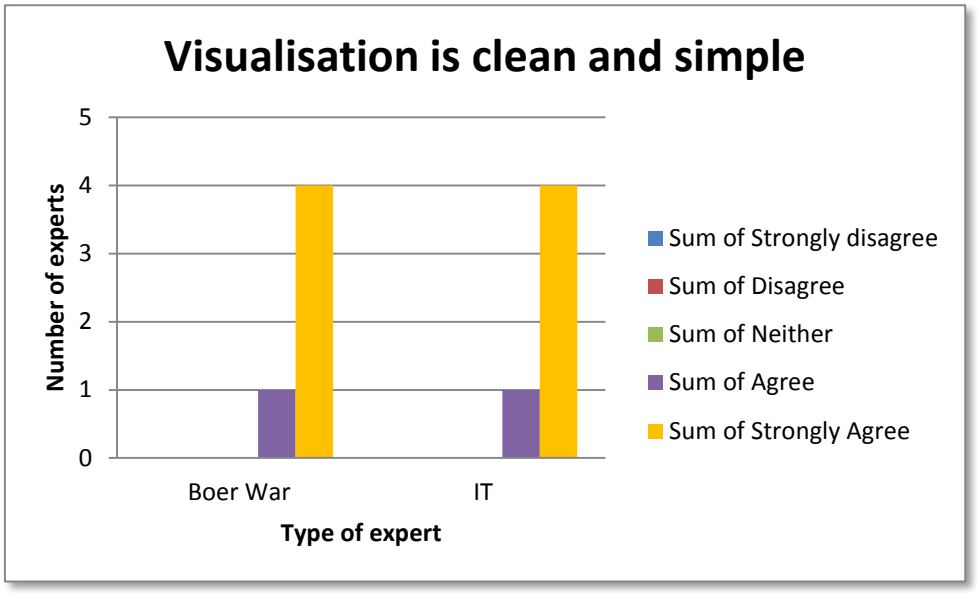


Figure 5.11: Visualisation is clean and simple

In general, 80% of the experts strongly agreed with the statement, while 20% agreed. Thus, it could be concluded that the visualisation was clean and simple.

8. Testimonies/documents and facts are easily distinguishable from each other.

In order to evaluate whether testimonies/documents and facts were easily distinguishable from each other, responses to the following statement were considered:

- *Testimonies/documents and facts are easily distinguishable from each other*.

As indicated in the pivot table (Table 5.14) and illustrated in the bar graph (Figure 5.12) on the following page, all the Anglo-Boer War experts strongly agreed with the statement. Four IT experts strongly agreed and one IT expert indicated that he or she neither agreed nor disagreed with the statement.

Table 5.14: Testimonies/documents and facts are easily distinguishable from each other

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	0	5
IT	0	0	1	0	4
Grand Total	**0**	**0**	**1**	**0**	**9**

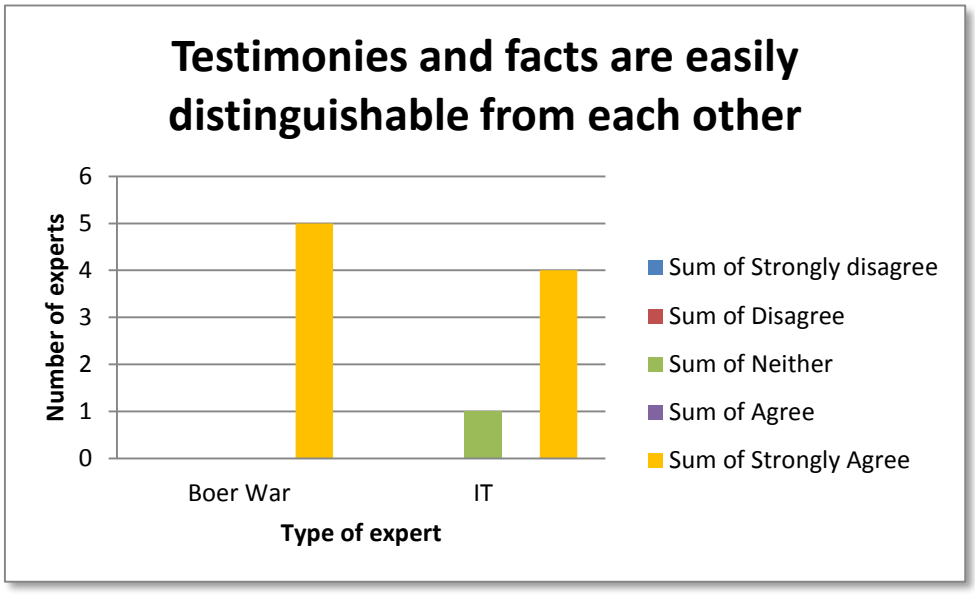


Figure 5.12: Testimonies and facts are easily distinguishable from each other

In general, 90% of the experts strongly agreed with the statement. Thus, it could be concluded that the testimonies or documents and facts were easily distinguishable from each other.

9. The visualisation leads to misleading interpretations.

In order to evaluate whether the visualisation could lead to misleading interpretations, responses to the following statement were considered:

- *The visualisation leads to misleading interpretations.*

As indicated in the pivot table (Table 5.15) and illustrated in the bar graph (Figure 5.13) on the following page, three Anglo-Boer War experts strongly disagreed and two Anglo-Boer War experts disagreed with the statement. Three IT experts strongly disagreed, one IT expert disagreed and one IT Expert indicated that he or she neither agreed nor disagreed with the statement.

Table 5.15: The visualisation leads to misleading interpretations

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	3	2	0	0	0
IT	3	1	1	0	0
Grand Total	**6**	**3**	**1**	**0**	**0**

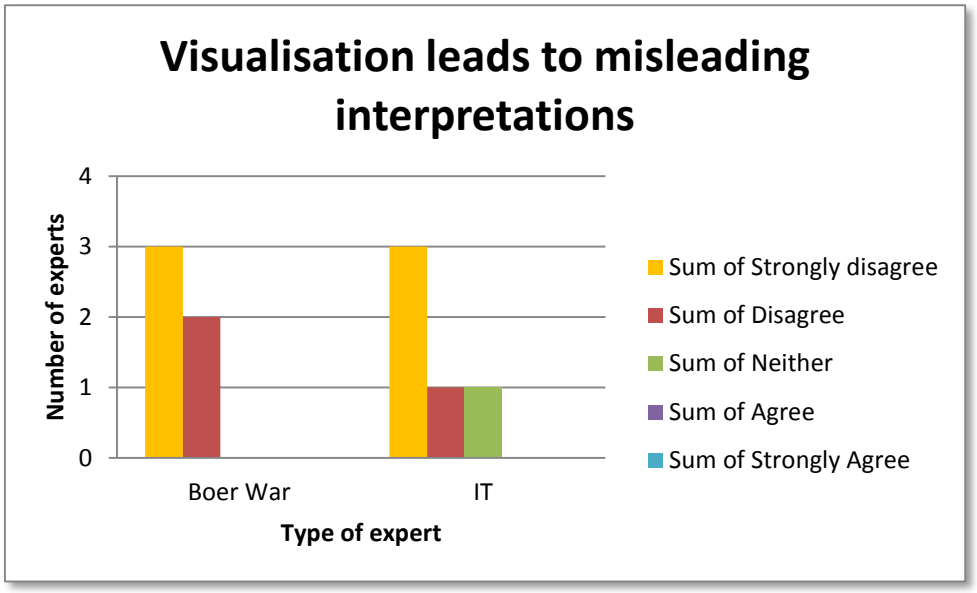


Figure 5.13: Visualisation leads to misleading interpretations

In general, 60% of the experts strongly disagreed with the statement, while 30% disagreed, which indicates that a total of 90% of respondents disagreed with the statement. Thus, it could be concluded that the visualisation did **NOT** lead to misleading interpretations.

10. The visualisation is interpretable.

In order to evaluate whether the visualisation was interpretable, responses to the following statement were considered:

- *The visualisation is interpretable.*

As indicated in the pivot table (Table 5.16) and illustrated in the bar graph (Figure 5.14) on the following page, four Anglo-Boer War experts strongly agreed and one Anglo-Boer War expert agreed with the statement. Four IT experts agreed and one IT expert strongly agreed with the statement.

Table 5.16: The visualisation is interpretable

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	1	4
IT	0	0	0	4	1
Grand Total	**0**	**0**	**0**	**5**	**5**

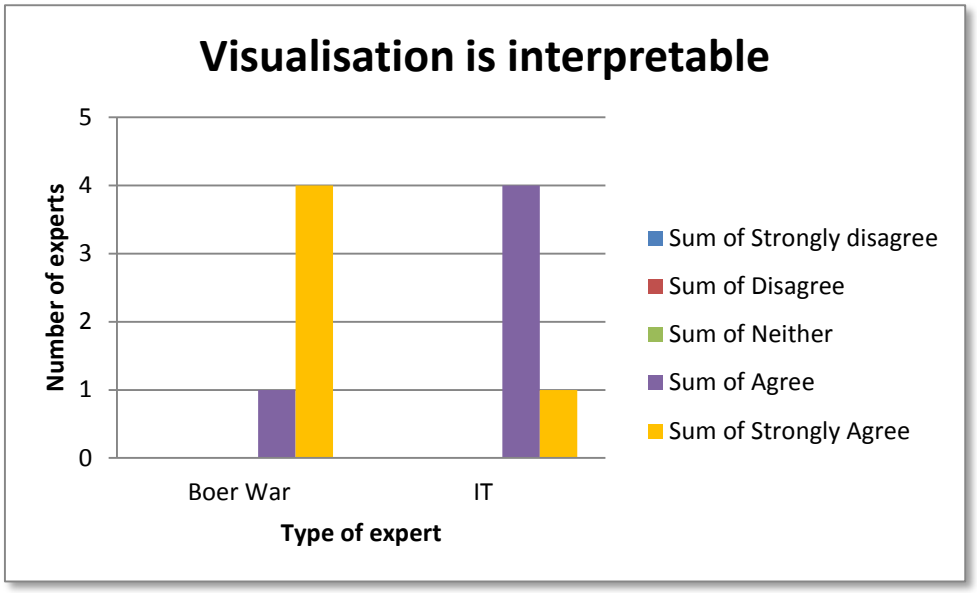


Figure 5.14: Visualisation is interpretable

In general, 50% of the experts strongly agreed with the statement, while 50% agreed. Thus, it could be concluded that the visualisation was interpretable.

In this section, feedback received from the demonstration participants was analysed in order to determine whether design principles and guidelines were applied to the design of the visual representation artifact. Except for one analysed feedback statement, the feedback received from the demonstration indicated that design principles and guidelines were applied to the design of the visual representation artifact. It could not be concluded with certainty that it was not necessary to add additional functionality to show less testimonies and facts. Future research could therefore consider additional functionality to show less testimonies and facts.

The feedback statement with regard to the usefulness of the solution artifact is analysed in the next section.

5.1.2.3 Feedback statement with regard to usefulness of the solution artifact

The feedback statement with regard to the usefulness of the solution artifact is analysed in this section.

1. The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history.

In order to evaluate whether the visualisation was useful for obtaining insights into aspects of the Anglo-Boer War history, responses to the following statement were considered:

- *The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history.*

As indicated in the pivot table (Table 5.17) and illustrated in the bar graph (Figure 5.15, all the Anglo-Boer War experts strongly agreed with the statement. One IT expert agreed, three IT experts strongly agreed and one IT expert indicated that he or she neither agreed nor disagreed with the statement.

Table 5.17: The visualisation is useful for obtaining insights into aspects of the Anglo-Boer War history

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Boer War	0	0	0	0	5
IT	0	0	1	1	3
Grand Total	**0**	**0**	**1**	**1**	**8**

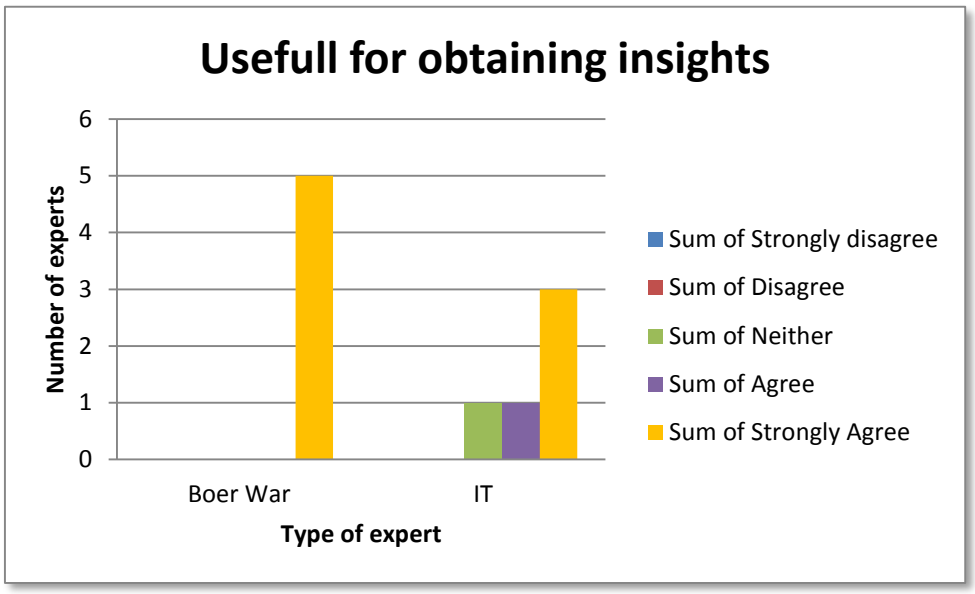


Figure 5.15: Useful for obtaining insights

In general, 80% of the experts strongly agreed with the statement, while 10% agreed. Thus, it could be concluded that the visualisation was useful for obtaining insights into aspects of the Anglo-Boer War history.

As part of the feedback questionnaire, the participants could also add comments and recommendations. The feedback comments and recommendation are discussed in the next section.

5.1.2.4 Feedback comments and recommendations

Comments and recommendations received from either the feedback questionnaires or during the demonstration sections are briefly discussed in this section. The following comments and recommendations were received:

- “It will be of great value for researchers of the Anglo-Boer War.” (Translated from Afrikaans)
- “As researcher, I was greatly impressed with this method and I believe that it will be of great value. The method opens up the Huyser collection for researchers. I have confidence that this study can be used by other researchers.” (Translated from Afrikaans)
- “An excellent visualisation method – I was very impressed.” (Translated from Afrikaans)
- “Very good aid. Interesting.” (Translated from Afrikaans)
- “The option to see which facts are the most coherent would be good.”
- “I would consider using both the translated text as well as the original (Dutch/Afrikaans) text in the computer program. I also believe that this functionality could be useful when including different sets of information, for example, comparing the Huyser collection with information included in a different but relevant collection (which also includes the stories of women from the ABW concentration camps). I believe this form of visualisation could be very useful for research purposes.”
- “Excellent way of visualising a data set.”
- “Very useful to visualise the coherence/opposite facts.”

The following observations were also made during the demonstration sessions:

- The use of the solution to visualise the coherence of facts describing other research domains such as medical research was identified.
- The use of the solution to assist qualitative research was identified.
- The ability to sort the facts not just alphabetically, but by the number of occurrence in the document set was identified.

- The functionality compare records contained not just one textual record set, but also different records set; therefore, adding more validity to the described facts.

- A requirement exists to have the solution made available to general users either commercially or as open source software.

- The purpose of the visualisation and the interpretation of the visual elements in the visualisation was clear to the users from the start.

As a result of the evaluation of the feedback given by the demonstration participants, it is deduced that:

- It is evident from the evaluation of the feedback received for Statements 6 and 7 discussed in Section 5.1.2.1, as well as in the comments received from the participants, that the instantiated solution contributed towards:

    - identifying components of a process to extract facts from a textual document set and to transform the unstructured data into a structured form;
    - identifying components of a process to visually represent the coherence of facts in a textual.

- Design principles and guidelines were considered in the design of the solution as evident in the evaluation of the feedback received from the feedback statements as discussed in Section 5.1.2.2. This is also evident from comments received from the participants.

- The instantiated solution is useful as is evident in the evaluation of the feedback received from respondents as discussed in Section 5.1.2.3. This is also evident from comments received from the participants.

Resulting from the feedback received from the questionnaire, it is suggested that the following enhancements and improvements be made to the solution:

- Consider additional functionality to show less testimonies and facts.

- Consider adding functionality to display the original untranslated source text to the user.

- Consider adding functionality allowing the user to also display the coherence of facts across different source sets.

- Consider functionality allowing the user to select how the facts in the dropdown lists should be sorted – either alphabetically or by the number of occurrences in the document set.

- Consider making the solution available to users across different domains. This involves a consideration of making the integration between the information extraction and visual representation artifacts seamless. It includes making the solution available to general users either commercially or as open source software.

Having discussed the development, demonstration and evaluation of the instantiated *final* specific solution for the visual representation of the coherence of facts in a textual document set, the *final* model which was informed by the instantiated solution is discussed next.

5.2 THE FINAL MODEL

The *final* model for visually representing the coherence of facts in a textual document set, is presented in this section. The resulting model is depicted in Figure 5.16 on the following page. The model was built as an IT using three phases of design science research (Figure 1.3.

On inspecting the model depicted in Figure 5.16 on the following page, it is suggested that the model fits in with the working definition of Information Visualisation as provided in Section 4.1.2.3 as follows:

> *The transformation from lower-level data to visual representation of meanings extracted from the data, where the extraction is either a computational process or a human transcription process; the aim is to explore data and create new insights.*

The proposed model is discussed next by referring to each component in the model:

- *The textual source set* consists of the source containing the facts that will be extracted and of which the coherence will be visually represented in subsequent processes. The source set can include handwritten words, digital textual documents or oral story tellings (in the case of indigenous knowledge).

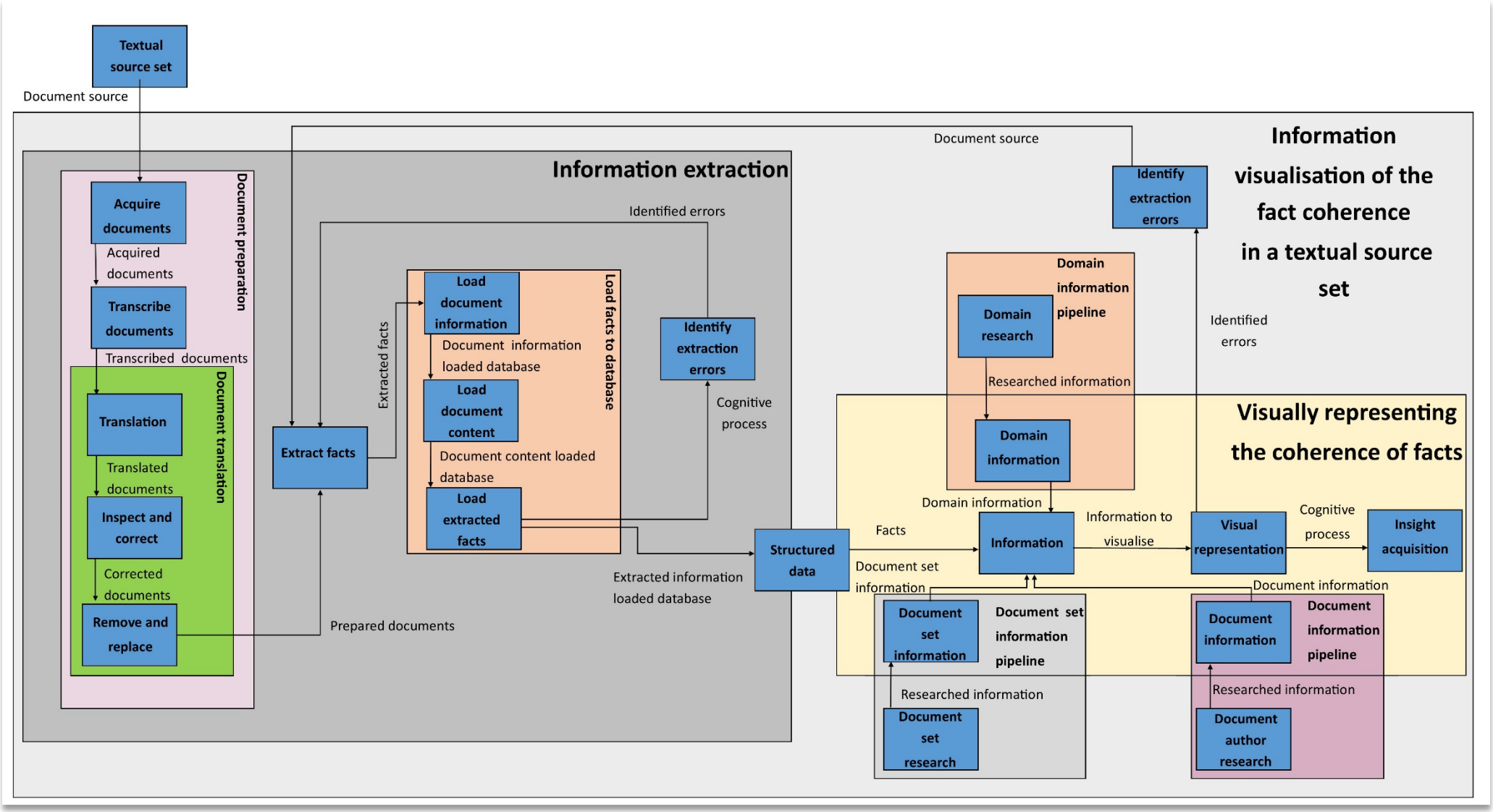


Figure 5.16: Final model - resulting Information Visualisation model

- *The information extraction* process prepares the textual documents, extracts the facts and loads the extracted facts into the database as follows:
    - *Document preparation.* Before the facts can be extracted from the textual source, the source set needs to be prepared. The document preparation process includes the acquiring, transcribing and translation of the documents as follows:
        - *Acquire documents.* The documents need to be acquired from the source. The acquired documents can include the photographic copy of handwritten words, copying of digital textual documents or the capture (into textual records) of oral story tellings (in the case of Indigenous Knowledge).
        - *Transcribe documents.* In cases where the textual documents are not in a digital format (like handwritten words), the documents need to be transcribed.
        - *Document translation.* In order to reach a larger audience, the textual documents need to be translated into a language understandable by most people. English is proposed as a possible language to translate the document into. The translation, in addition, involves the removing and replacing of defamatory words and names of protected individuals.
            - *Translation.* In order to assist in translating the textual documents, a translation tool such as *Google translator* is suggested.
            - *Inspect and correct.* Using a language translator is not entirely reliable and therefore the translated textual record needs to be inspected for spelling and grammar errors, and where needed, corrections should be made.
            - *Remove and replace.* In order to remain ethical, defamatory words should be removed and replaced with acceptable words. The identities of certain persons might have to be protected and therefore identities can be protected by replacing names of certain individuals with coded names.
    - *Extract facts.* For the purpose of the research study, a manual extraction method was described in Section 3.1.2.1. For the purpose of giving the reader an idea of what the manual information extraction involves, the

process of using ATLAS.ti to extract facts is discussed in Section 3.2.2. It is proposed that employing unsupervised and semi-supervised methods for extracting facts from the textual source set can be the focus of future research.

- *Load facts into database.* The facts extracted are loaded into the database system in the following three steps:
  - *Load document information.* Information identifying each document is loaded into the database by means of a simple web service that was developed as part of the research.
  - *Load document content.* Having loaded information identifying each document, the textual content of the document is loaded next. The document content is loaded by means of a small software program written in the PHP programming language that was developed as part of the research.
  - *Load extracted facts.* All the facts that were extracted are loaded next into the database by means of a simple web service that was developed as part of the research.
- *Identify extraction errors.* Identification of possible errors (mainly due to human error) can be done at two different points in the process. The errors can be identified by means of an inspection of the extracted facts, or by using the visual representation to visually highlight possible errors.
- *Structured data.* The output of the information extraction and database loading activities is a structured data set.

- *Visually representing the coherence of facts.* The visual representation of the coherence of facts requires that the information contained in structured data, domain information, document information and document set information must be visually represented in order to assist the user in acquiring insight into the coherence of the facts in the document set. Ultimately, the user should be able to make an assumption on the validity of facts contained in the record set.
  - *Information.* The information to be visually represented is fed by the extracted facts (in a structured data format), domain information, document information and the document set information.
  - *Domain information pipeline.* In order to assist the user with the insight-acquiring process, researched domain information is fed to the visual

representation. The process of feeding the domain information to the system consists of the following:

- *Domain research.* This consists of an external research process to collect domain information on the extracted facts.
- *Domain information management.* In order for the visual representation to display correct domain information, the collected domain information needs to be added, deleted or updated.

o *Document source information pipeline.* In order to assist the user with the insight-acquiring process, researched information with regard to the individual documents is fed to the visual representation. The process of feeding the document information to the system consists of the following:

- *Document author research.* This consists of an external research process to collect author information for each document.
- *Document information management.* In order for the visual representation to display correct document information, the collected document information needs to be added, deleted or updated.

o *Document set information pipeline.* In order to assist the user with the insight-acquiring process, researched information with regard to the document set is fed to the visual representation. The process of feeding the document set information to the system consists of the following:

- *Document set research.* This consists of an external research process to collect information about the document set.
- *Document set information management.* In order for the visual representation to display correct document set information, the collected document set information needs to be added, deleted or updated.

o *Visual representation.* The visual representation of the coherence of facts is presented to the user to assist the user in acquiring insight into the possible truth value of extracted facts.

In order to support the cognitive process to make an assumption on the possible truth value of a fact, the user is visually presented with the coherence of a selected fact in the document set, domain details (on

demand) on a selected fact, the context in which the selected fact appears in a document, document details (on demand) containing an occurrence of the selected fact and details of the document set.

- *Insight acquisition.* According to Ackoff (1989), human understanding is a cognitive process that is built upon layers (data, information, knowledge, understanding and wisdom). The information visualisation model developed by the research study accommodates this cognitive process by supporting these layers of human understanding.

  The *data* layer is provided by the textual records. Extracted facts provide the *information* layer. The *knowledge* layer is provided by the visual representation, which allows the user to acquire some knowledge on the coherence of a fact in a record set. Having an *understanding* that the coherence of a fact (according to the coherence theory of truth) is a possible indication that the fact might contain truth, ultimately leads to some domain *wisdom* with regard to the domain being described by the data.

Having described the developed of the model for visually representing the coherence of facts in a textual document set, the next two sections describe the process of demonstrating and evaluating the model.

5.2.1 FINAL MODEL DEMONSTRATION

In order to ensure research rigour, the final model of the research study (described in the previous section) was demonstrated and evaluated. This section of the dissertation describes the demonstration activity of the resulting model.

For the purpose of the research study, the model for visually representing the coherence of facts in a textual document set, was demonstrated by means of five individual interviews each having a duration of about one hour. During each interview, the components of the model were briefly described to the participants, which consisted of the same IT experts that were used in the demonstration of the instantiated specific software solution (see Table 5.1).

The decision to use five IT experts is based on Nielsen and Landauer's (1993) proposition that between three and five expert users are adequate for conducting user evaluations. After each demonstration interview, the participants were presented with a demonstration feedback questionnaire containing the information and statements listed in Table 5.18 on the following page.

Table 5.18: Model demonstration feedback questionnaire

Background

My name is Louis Engelbrecht, and as part of my Master of Computer Science degree at the University of South Africa, a software solution artifact was developed.
The aim of the software artifact is to visually represent the domain facts contained in a set of textual documents. For the purpose of this research, the Huyser collection of Anglo-Boer War testimonies is used as source for the visualisation.
After being provided with a demonstration of the resulting model developed by the research, you will be requested to evaluate the model by completing this form.

Participant information

1. What is your expertise in IT?

_____

_____

_____

_____

_____

2. What is your level of IT expertise?

_____

Instructions

1. Please complete the next section by marking your choice with an "X" in the box provided.
2. Please answer all questions.
3. Please note that each answer only requires a single choice.

Evaluation form

Please read the statements with regard to the demonstrated software below and how strongly you agree or disagree. Please use the following rating system:

Strongly disagree	Somewhat disagree	Neither agree or disagree	Agree	Strongly agree
1	2	3	4	5

Number	Statement	Rating				
1.	The model represents components for extracting information from unstructured textual records into a structured format and the application of a visualisation technique to the transformed set, in order to gain insight into the coherence of facts contained in the document set.	1	2	3	4	5
2.	Subcomponents for extracting information from unstructured textual records into a structured format are represented.	1	2	3	4	5
3.	Subcomponents for the application of a visualisation technique onto a structured data set, in order to gain insight into the coherence of facts contained in the document set, are represented.	1	2	3	4	5
4.	The represented model represents the components of the visualisation solution that was previously demonstrated.	1	2	3	4	5

For each statement, the participant had to indicate how strongly he or she agreed or disagreed by selecting a value from one to five according to a five-point Likert scale. The statements presented to the participants were carefully defined in order to be used

as input for the evaluation of the model as discussed in Section 5.2.2. Table 5.19 contains a list of the feedback questionnaire statements, each with an indication of its contribution to the model evaluation.

Table 5.19: Contributions of feedback statements to model validation

Statement	Evaluation contribution
1. The model represents components for extracting information from unstructured textual records into a structured format and the application of a visualisation technique to the transformed set, in order to gain insight into the coherence of facts contained in the document set.	Evaluate whether the model achieves the purpose of the study, namely to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set.
2. Subcomponents for extracting information from unstructured textual records into a structured format are represented.	Evaluate whether the model contains the components of the *initial* information extraction model.
3. Subcomponents for the application of a visualisation technique to a structured data set, in order to gain insight into the coherence of facts contained in the document set, are represented.	Evaluate whether the model contains the components of the *intermediate* visual representation model.
4. The represented model represents the components of the visualisation solution that was previously demonstrated.	Evaluate whether the model represents the components of the instantiated software solution represented to the participants, i.e. did the specific software solution inform the development of the model?

The result from the feedback questionnaire is used as input into the evaluation of the resulting model as discussed in the next section.

5.2.2 FINAL MODEL VALIDATION

The model for visually representing the coherence of facts in a textual document set produced by the research study is evaluated in this section.

The evaluation is guided by feedback received from IT experts after being presented with the resulting model. The feedback is in the form of a five-point Likert scale questionnaire that was presented to each participant after the model was demonstrated during interviews. The data is analysed using *Microsoft Excel.* To analyse the data, the researcher used pivot tables and charts for each of the statements presented in the feedback questionnaire.

The statements presented in the feedback questionnaire were designed in order to determine the following:

- Whether the model achieves the purpose of the study, namely to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set?

- Whether the model contains the components of the *initial* information extraction model?

- Whether the model contains the components of the *intermediate* visual representation model?

- Whether the model represents the components of the instantiated software solution represented to the participants, i.e. did the specific software solution inform the development of the model?

The next section analyses the feedback data in order to determine whether the model contains components to extract facts and transform unstructured textual data into a structured form.

5.2.2.1 Was the purpose of the study achieved in the model?

In order to evaluate whether purpose of the study was achieved in the model, the responses to the first statement were considered as follows:

> *The model represents components for extracting information from unstructured textual records into a structured format and the application of a visualisation technique to the transformed set, in order to gain insight into the coherence of facts contained in the document set.*

As indicated in the pivot table (Table 5.20) and illustrated in the bar graph (Figure 5.17) on the following page, one IT expert agreed and four IT experts strongly agreed.

In general, 80% of the experts strongly agreed with the statement, while 20% agreed. Thus, it could be concluded that the research purpose was achieved in the model.

Table 5.20: Purpose of the study achieved in the model

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
IT	0	0	0	1	4
Grand Total	**0**	**0**	**0**	**1**	**4**

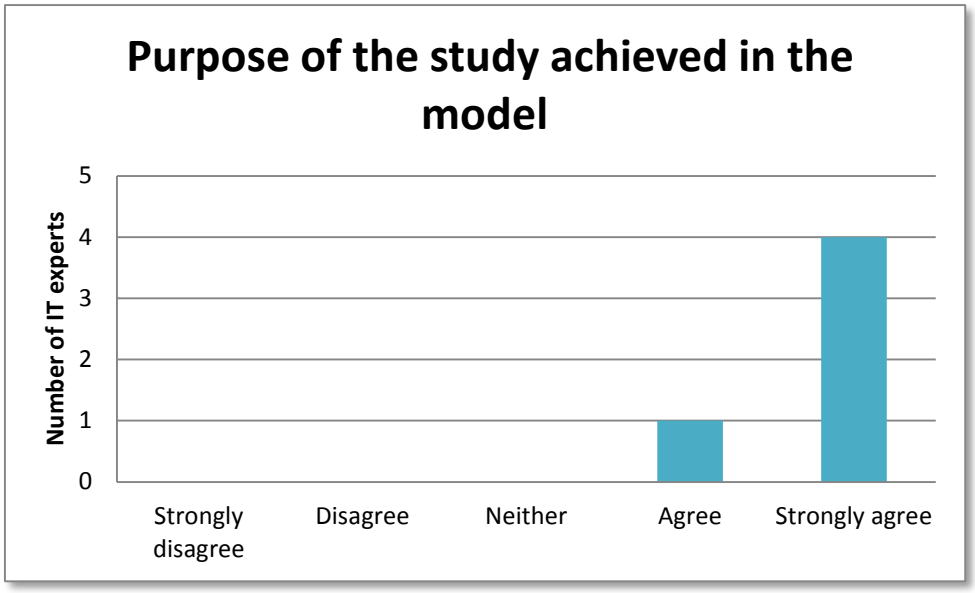


Figure 5.17: Purpose of the study achieved in the model

Having concluded that the research purpose was achieved in the model, the next section evaluates whether model contains components to extract information from unstructured records.

5.2.2.2 Does the model contains components to extract information from unstructured records?

In order to evaluate whether the model contains components to extract facts and transform unstructured textual data into a structured form, the responses to the second statement were considered as follows:

> *Subcomponents for extracting information from unstructured textual records into a structured format are represented.*

As indicated in the pivot table (Table 5.21) and illustrated in the bar graph (Figure 5.18) on the following page, two IT experts agreed, two IT experts strongly agreed and one IT expert indicated that he or she neither agreed nor disagreed with the statement.

In general, 40% of the experts strongly agreed with the statement, while 40% agreed. Thus, it could be concluded that the objective to extract information from textual sources was achieved in the model.

Table 5.21: Model contains information extraction components

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
IT	0	0	1	2	2
Grand Total	**0**	**0**	**1**	**2**	**2**

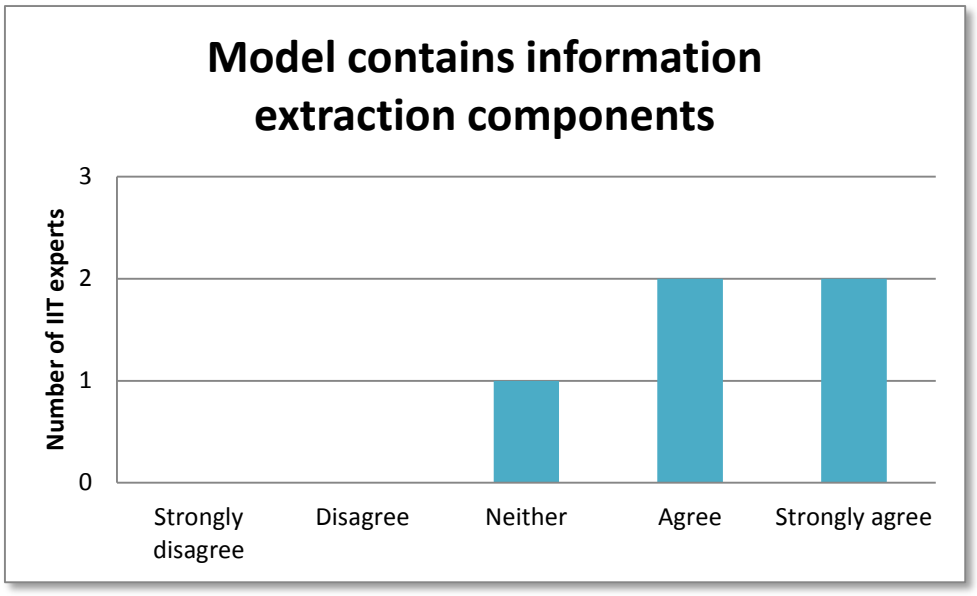


Figure 5.18: Model contains information extraction components

Having concluded that the model contains information extraction components, the next section evaluates whether the model contains components to visually represent the coherence of facts in a structured data form.

5.2.2.3 Does the model contains visual representation components?

In order to evaluate whether the model contains components to visually represent the coherence of facts in a structured data form, the responses to the third statement were considered as follows:

> *Subcomponents for the application of a visualisation technique onto a structured data set, in order to gain insight into the coherence of facts contained in the document set, are represented.*

As indicated in the pivot table (Table 5.22) and illustrated in the bar graph (Figure 5.19) on the following page, two IT experts agreed, two IT experts strongly agreed and one IT expert indicated that he or she neither agreed nor disagreed with the statement.

In general, 40% of the experts strongly agreed with the statement, while 40% agreed. Thus, it could be concluded that the objective to visually represent the coherence of facts was achieved in the model.

Table 5.22: Model contains visual representation components

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
IT	0	0	1	2	2
Grand Total	**0**	**0**	**1**	**2**	**2**

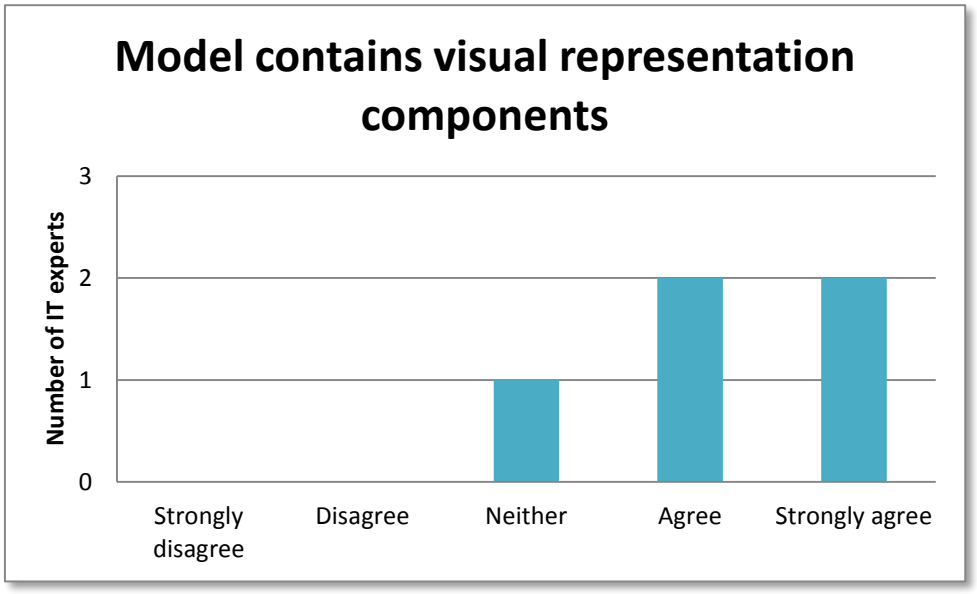


Figure 5.19: Model contains visual representation components

Having concluded that the model contains components to visually represent the coherence of facts in a structured data form, the next section evaluates whether the model represents the components of the instantiated software solution represented to the participants, i.e. did the specific software solution inform the development of the model?

5.2.2.4 Was the development of the model informed by the construction of the instantiated software solution?

In order to evaluate whether the model was informed by the construction of the instantiated software solution, the responses to the fourth statement were considered as follows:

> *The represented model represents the components of the visualisation solution that was previously demonstrated.*

As indicated in the pivot table (Table 5.23) and illustrated in the bar graph (Figure 5.20), one IT expert agreed and four IT experts strongly agreed.

In general, 80% of the experts strongly agreed with the statement, while 20% agreed. Thus, it could be concluded that the resulting model was informed by the construction of the instantiated software solution presented to the demonstration participants.

Table 5.23: Model informed by instantiton of the software solution

Row Labels	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
IT	0	0	1	2	2
Grand Total	**0**	**0**	**1**	**2**	**2**

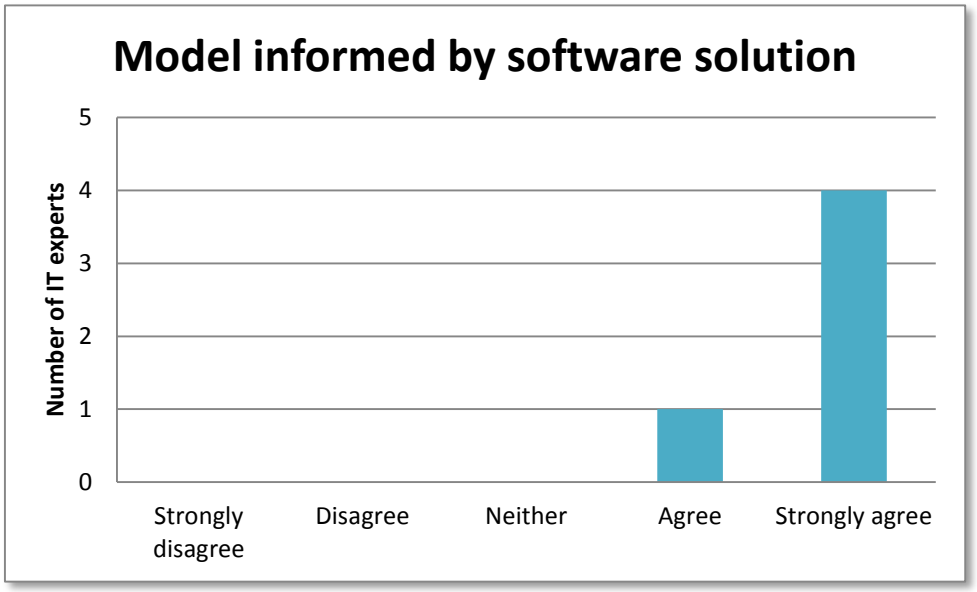


Figure 5.20: Model informed by instantiton of the software solution

In this chapter, a model for visually representing the coherence of facts in a textual document set, was presented. The resulting model was demonstrated to five IT experts and an evaluation based on a feedback questionnaire was conducted. As a result of the evaluation, it is concluded that:

- The model achieves the purpose of the study, namely to develop a model to transform unstructured textual data and to visually represent the coherence of facts in the set.
- The model contains the components of the *initial* information extraction model.
- The model contains the components of the *intermediate* visual representation model.
- The model represents the components of the instantiated software solution represented to the participants, i.e. did the specific software solution inform the development of the model.

Having presented the model for visually representing the coherence of facts in a textual document set, the dissertation is concluded in the next chapter.

CHAPTER 6: CONCLUSION

This chapter concludes the research study by presenting the results of the research. The research questions of the study are referenced in this chapter in order to show that they are answered by the research.

A summary of the chapters of the dissertation is presented in Section 6.1. An overview of the study as well as a review of the answers provided to the research questions is discussed in Section 6.2. Contributions made by the study is discussed in Section 6.3 and Section 6.4 reviews the delineation and assumptions of the research study. Section 6.5 presents a scientific, methodologic and personal reflections on the study. Possible future research is discussed in Section 6.6, which is followed by the final concluding remarks in Section 6.7.

6.1 SUMMARY OF CHAPTERS

In this section, a brief summary of the dissertation chapters is given.

- *Chapter 1. Introduction* defined the study by providing the research background, research purpose and research questions of the study. In addition, the importance of the research, delineations and assumptions as well as ethical considerations of the study is discussed.

- *Chapter 2*. *Research design and methodology* presented the main research methodology followed in order to address the purpose of the research and the research questions. In essence a design science research approach is followed by using an adaption of the design science research process model from Peffers *et al*. (2007).

- *Chapter 3*. *Development of the initial model – textual information extraction* describes the design and development of the *initial* model to extract information from textual documents.

- *Chapter 4*. *Development of the intermediate model – visual representation* described the design and development of the *intermediate* model to visually represent the coherence of facts contained in a structured data form.

- *Chapter 5*. *Demonstration and evaluation of the specific software solution and validation of the final model* described how to specific software that informed the realisation of the final *model* was demonstrated and evaluated. The chapter in addition reviewed the validation of the *final* model.

- *Chapter 6*. *Conclusion* presents the results of the study.

The overview of the study is presented next in Section 6.2

6.2 OVERVIEW OF THE STUDY

An overview of the study is presented in this section. The process followed to answer the research question (Section 6.2.1) and a reflection on the key findings of the research (Section 6.2.2) is provided in this section.

6.2.1 The research process

In the development of the NIKMAS system for the South African Department of Science and Technology (DST), it soon became apparent that Indigenous Knowledge being described by different traditional communities was in coherence. It is proposed that having insight into the coherence of the facts (for instance, using the same plants for the same illnesses and the employment of similar preparation methods) contained in the Indigenous Knowledge description would be a step closer to making an assumption about the validity of the knowledge.

The assumption that the coherence of facts contained in the different Indigenous Knowledge descriptions is an indication of its possible truth value is based on the coherence theory of truth. In order to assist the process of acquiring insight into the coherence of facts (Indigenous Knowledge), the research study proposed that a visual representation of the coherence of the Indigenous Knowledge facts could assist the insight-acquiring process. However, the Indigenous Knowledge exists orally and in the customs of the traditional communities. Therefore, the knowledge needs to be acquired, captured into text and subsequently the facts from the textual document set needs to be extracted and loaded into a database. Only after the Indigenous Knowledge has been transformed into a structured form, can it be visualised.

The *main research question* was therefore defined as:

> *What components can be combined into a model to transform unstructured textual data and to visually represent the coherence of facts contained in the set?*

In order to answer the main research question, the following *subordinate research questions* were formulated:

1. What components can be combined into a model to extract facts and transform unstructured textual data into a structured form?

2. What Visualisation design principles can be applied to guide the design of an Information Visualisation interface?

3. What components can be combined into a model to visually represent the coherence of facts contained in a structured data form?

In order to answer subordinate research questions 1 and 3, a design science research approach was followed, which is discussed in Chapter 2. With regard to answer research question 2, Visualisation design principles were identified by means of a literature review as discussed in Section 4.1.3. A summary of the research process followed in the study is provided in Table 6.1 on the following page.

The research process is summarised as follows: A model artifact for the visual representation of the coherence of facts in a textual document set were constructed mainly through three phases of design science research. These phases consisted of:

1. Phase 1. An *initial* information extraction phase.

2. Phase 2. An *intermediate* visual representation phase.

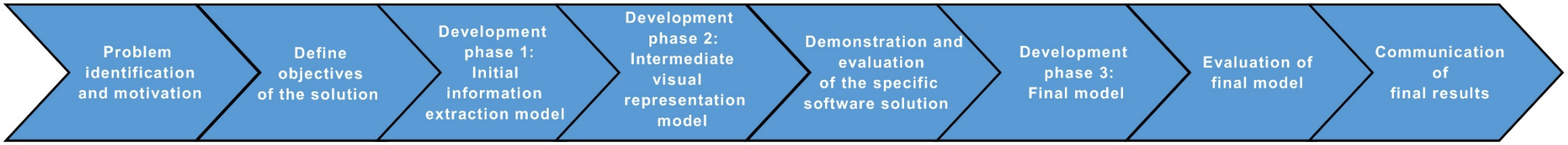
3. Phase 3. *The final* model which was created by combining the results of the *initial* and *intermediate* phases.

A reflection on the key finding is discussed next.

6.2.2 Reflection on key findings

The findings of the research were derived from a literature review as well as from the development and instantiation of a specific software solution. The specific software solution consisted of two parts, namely an *initial* information extraction and an *intermediate* visual representation part.

Table 6.1: Summary of the research process



Section 1.2 & 1.4	Section 1.2.4	Chapter 3	Chapter 4	Section 5.1	Section 5.3	Section 5.2	Chapter 6
The starting point for the research was the identification of the research problem and the motivation of the importance of having a visual representation of the coherence of facts in a textual record set, especially where the textual record set describes indigenous	Research objectives for achieving the purpose of the research were identified during this task. At the end of the study, the researcher have to reflect on these objectives in order to determine whether the purpose of the study was achieved	A specific software solution for the extraction of information from textual records was developed during this task. The specific software solution informed the construction of a model to extract information from textual records.	A specific software solution for the visual representation of the coherence of facts in a data set was developed during this task. The software solution informed the construction of a model to visually represent the coherence of facts in a data set.	The specific software solution that was developed in parts during development phase 1 and 2 was demonstrated to five IT experts as well five experts in the domain of the information (Anglo-Boer War) being described. Feedback were used for evaluation	A final model for the visual representation of the coherence of facts in a textual document set was developed from the instantiation of a specific software solution consisting of 2 software parts namely an information extraction and a visual representation part.	The final model was presented to 5 IT experts. Feedback was received from them in the form of a questionnaire. The feedback was then used to evaluate whether the purpose of the research achieved in the model.	The final results of the research was communicated in the conclusion chapter of this dissertation.
Output	**Output**	**Output**	**Output**	**Output**	**Output**	**Output**	**Output**
As a result, the purpose of the research, the main research question and the sub-research questions of the study were identified.	The research objectives required to achieve the purpose of the study.	A model that answers research question 1, namely a model for extracting information from textual records.	A model for visually representing the coherence of facts and a set of visualisation design principles.	A reflection on whether the specific software solution contributed to the achievements of the research.	A model to transform unstructured textual data and to visually represent the coherence of facts contained in the set	An evaluation of the final model.	Conclusion chapter of the dissertation containing the final results of the research study.

Subordinate research question 1

What components can be combined into a model to extract facts and transform unstructured textual data into a structured form?

Subordinate research question 1 was addressed in Chapter 3 of the dissertation. A design science research approach was followed to create an *initial* specific software solution for extracting information from textual records. The process of creating the specific information extraction solution identified components which were combined into an *initial* model (depicted in Figure 3.11). Validity of the model resides in the knowledge that the model is based on an instantiation that was demonstrated and evaluated (discussed in Chapter 5).

The creation of the information extraction solution was informed by an initial literature review (Section 3.1) which, amongst other outputs, identified possible extraction methods and software tools. Supervised, semi-supervised and unsupervised methods where identified for extracting facts from textual sources. At the onset of the study it was envisaged that a semi-supervised or unsupervised method would be used. As the study progressed it was however realised that combining semi-supervised or unsupervised extraction research, with the intended visualisation of fact coherence would require an amount of work outside the scope of a master's degree study. The realisation was confirmed by an academic audience when the intended study was presented at both the SAICSIT 2014 Master's and Doctoral Symposium (SAICSIT 2014 Master's and Doctoral Symposium, 2014) and at the FASTAR/Espresso Workshop 2014 (Engelbrecht, 2014). As a result, the research study proposes that a possible reason for the gap between the research fields of Information Extraction and Information Visualisation (identified in Section 1.2.1) resides in the amount of research required by each.

Subordinate research question 2

What visualisation design principles can be applied to guide the design of an information visualisation interface?

Subordinate research question 2 was addressed in Section 4.1.3 as part of the review on literature relevant to the visualisation of information coherence. As a result a list of 32 information visualisation design principles as listed in Table 4.3 were compiled. These principles are not strict rules, but serve as guide to the design process of an Information Visualisation interface. The use of the principles is demonstrated in the design of the interface for the model required by subordinate research question 3 (see Section 4.2.1.1.5). In addition, Section 5.1 shows how the principles can be used a feedback questionnaire to determine whether design principles was incorporated into a visualisation interface design.

In addition to demonstrating that the principles can be used in a real visualisation design, value has been added through a process of peer review and academic scrutiny through a conference paper (Engelbrecht, Botha, & Alberts, 2014) and a journal article (Engelbrecht *et al*., 2015).

Subordinate research question 3

What components can be combined into a model to visually represent the coherence of facts contained in a structured data form?

Subordinate research question 3 was addressed in Chapter 4 of the dissertation. A design science research approach was followed to create an *intermediate* specific software solution for visually representing the coherence of facts in a structured data form.

The creation of the visual representation solution was informed by an initial literature review (Section 4.1) which, amongst other outputs, identified possible information visualisation tools and techniques that could be used in the design and creation of the user interface. From these a circular graph visualisation technique was considered as suitable for the intended purpose of the study. The D3.js software library was furthermore chosen to implement the visualisation design. The process of creating the *intermediate* visual representation solution identified components which were combined into an *intermediate* model (depicted in Figure 4.30). Validity of the model resides in the knowledge that the model is based on an instantiation that was demonstrated and evaluated (discussed in Chapter 5).

Main research question

What components can be combined into a model to transform unstructured textual data and to visually represent the coherence of facts contained in the set?

The main research question of the study addresses the problem of visually representing the coherence of facts contained in unstructured textual data; with intend to assist the process of making an assumption about the validity of information in the data. The three subordinate questions of the research contributed to addressing the main research question.

Subordinate question 1 provided an *initial* model for extracting information from unstructured textual data and transforming the data into a structured form. In addition the *initial* model provided by subordinate question 1 also addressed the situation where the data is not yet in a textual format, for instance when the data source is in an oral format.

Subordinate question 2 provided Information Visualisation design principles that could be used as basis for developing the visual representation model required in the next subordinate question. Subordinate question 3 provided an *intermediate* model for visually representing the coherence of facts in a structured data form

Subordinate questions 1 and 2 provide input to subordinate question 3. Combined together all three subordinate questions progress towards providing an answer to the main research question, namely a model for visually representing the coherence of facts in a textual document set. The model consists of 2 main components and various sub-components depicted in Figure 6.1 and listed in Table 6.2 on the following two pages.

Validity of the model resides in the knowledge that the model is based on an instantiation of both the *initial* information extraction and *intermediate* visual representation models which were demonstrated and evaluated (discussed in Chapter 5).

The significance and contribution of the research is discussed in the next section.

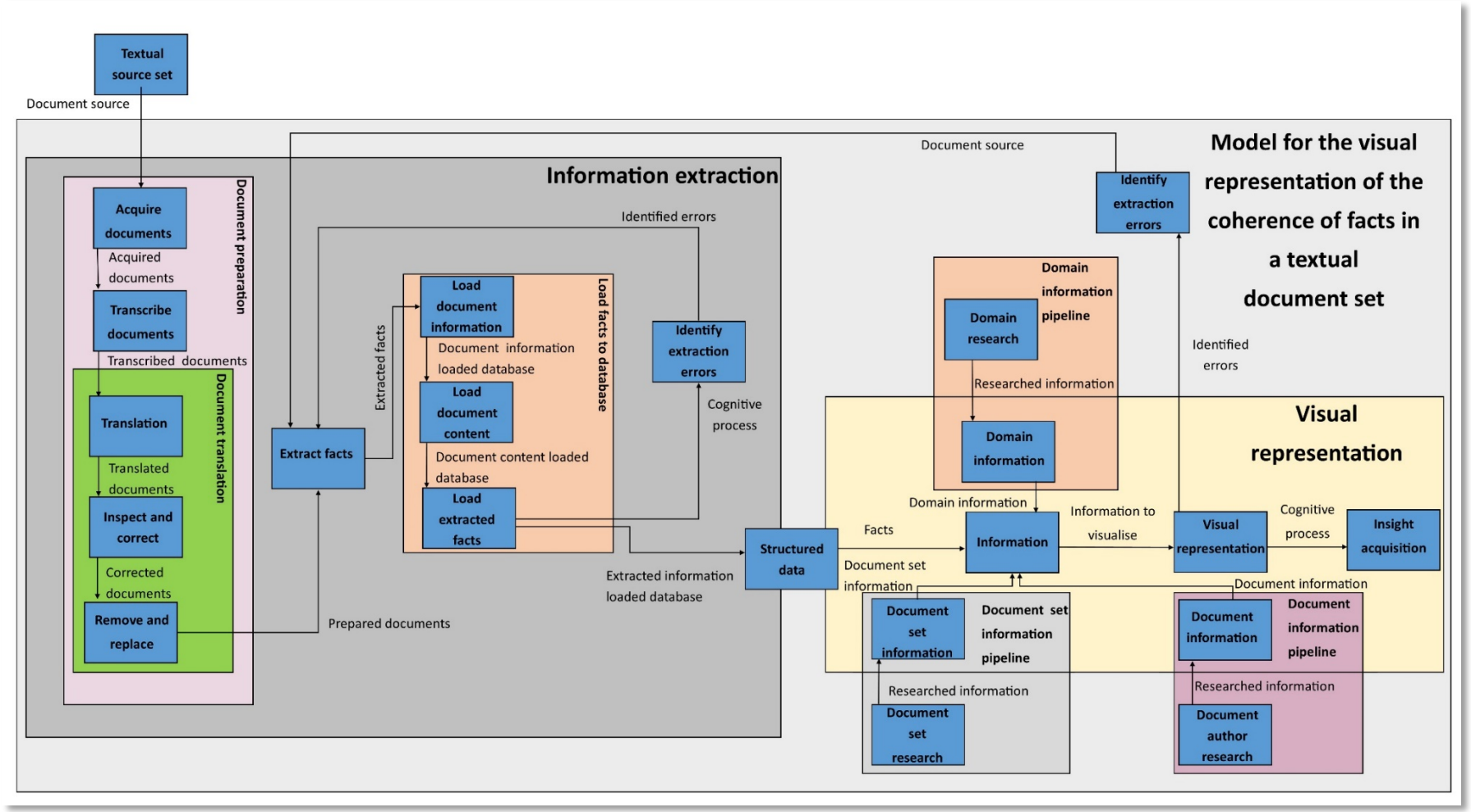


Figure 6.1: A model for visually representing the coherence of facts in a textual document set

Table 6.2: Model components for the visual representation of the coherence of facts in a textual document set

Main components	Secondary components	Tertiary components	Fourth level components
Information extraction	*Document preparation*	Acquire documents	
		Transcribe documents	
		Document translation	Translation
			Inspect and correct
			Remove and replace
	Extract facts		
	Load facts to database	Load document information	
		Load document content	
		Load extracted facts	
	Identify extraction errors		
	Structured data		
Visually representing the coherence of facts	*Structured data*		
	Information		
	Domain information pipeline	Domain research	
		Domain information management	
	Document source information pipeline	Document author research	
		Document information management	
	Document set information pipeline	Document set research	
		Document set information management	
	Visual representation		
	Insight acquisition		

6.3 SIGNIFICANCE AND CONTRIBUTION OF THE RESEARCH

The results of the study indicate that the study can make a contribution to the academic computer science community, especially in the research fields of Indigenous Knowledge, Information Extraction and Information Visualisation.

The significance of the research study and the contribution to the scientific body of knowledge is summarised as follows:

- A contribution is made to the study of Indigenous Knowledge by presenting a novel model for making an assumption on the validity of the facts contained in descriptions of indigenous knowledge.

- A contribution is made by providing a scientific evaluation method for making assumptions on Indigenous Knowledge and therefore contributing to the acknowledgement of Indigenous Knowledge containing truth value.

- A contribution is made to other fields of study such as History studies (testimonies of witnesses to a historical event), Medical studies (patient health records), Criminology (crime reports) and Political Science studies (political speeches) because it is proposed that the same Information Visualisation model can also be applied to other fields.

- A novel Visualisation method for presenting the coherence of facts in a structured data form is presented.

- A contribution is made to the study of Information Visualisation by adding derived definitions for Visualisation concepts to the scientific body of knowledge.

- An increased understanding of how information is visualised and a better grasp of how data, information, knowledge and wisdom relates to each other.

- A contribution is made to the study of Information Visualisation by providing a list of design principles and guidelines that can be referenced in future Information Visualisation projects.

- A better understanding of why there exists a gap between the study fields of Information Extraction and Visualisation is gained. During the study, it was realised that a large amount of research is required to combine unsupervised or semi-supervised extraction methods with the visualisation of the information. Therefore, a possible reason for the gap between the two research fields resides in the amount of research required by each.

In essence the knowledge gained from this study can guide other researchers to visually represent the coherence of facts in a textual document set, irrelevant of the domain being described the facts.

The delineation and assumptions of the research study is discussed in the next section.

6.4 DELINEATION AND ASSUMPTIONS OF THE RESEARCH STUDY

During the consideration of the research problem and the importance of the study, it became evident that for the purpose of the research, the scope of the study had to be set out and certain assumptions had to be made as follows:

- The proposed visualisation will only be applicable to a finite number of textual records not exceeding 50 documents. Larger sets will require future research to be done.

- Accessibility and usability issues were not taken into account.

- The research study proposed a suitable solution (both visual representation and information extraction solution) and not necessarily the most effective solution.

- The assumption is made that the existence of fact coherence in a textual document set is an indication of the possibility of truth contained in the fact. Although it is acknowledge that other theories of truth exist, they were not evaluated and compared to each other.

- Due to intellectual property issues and the availability of textual documents describing domain facts, Anglo-Boer War testimonies were used as data source. It is assumed that the proposed Information Visualisation model can be applied to both other study fields where textual records with similar characters as described in Section 1.2.1 exist.

- The work required to provide the most suitable Information Extraction method is considered as beyond the scope and amount of work required for a master's degree study. The study was therefore limited to providing a supervised, manual information extraction method. Whether a semi-supervised of unsupervised automatic method would be more efficient was therefore considered as out of scope for the study.

The listed delineation and assumptions of the study can be considered as guidelines for future areas of research. Having reflected on the delineation and assumptions of the research study, the next section provides a scientific, methodological and personal reflection of the study.

6.5 REFLECTIONS

Scientific, methodological and personal reflections on the study is provided in this section.

6.5.1 Scientific reflection

The study defined a model for the visual representation of the coherence of facts in a textual document set. The goal of the visual representation was to enable the user to make an assumption about the possible truth value of the facts. The requirement for such a model was identified during the development of the NIKMAS Indigenous Knowledge research project. Due to various reasons which include intellectual property restrictions, civilian war testimonies were used as source for the model. It is proposed that the model can be applied to both History and Indigenous Knowledge research domains since the data sets exhibit the same characteristics; both are in textual documents (or can be transformed into textual document) containing unstructured data which are secondary and open for scrutiny. The model therefore provides a scientific visualisation method that can be applied to any other domain where facts contained in textual documents are being scrutinised.

In addition to providing the Information Visualisation model, the study provides a set of Information Visualisation design principles and guidelines that can be used in future visualisation research, which can include an expansion of it.

By following a design science research approach during the research, the study provided a scientific case for design science research as being an appropriate approach to Information Visualisation research.

6.5.2 Methodological reflection

A methodological reflection on the study is provided next by considering various relevant questions.

- *Was the chosen research methodology the best for answering the research questions?*

In the process of identifying the components as required by both subordinate research questions 1 and 3, specific IT software solutions had to be designed and built. Furthermore, various authors state that design science research is an appropriate problem-solving paradigm for creating new innovations in the form of IT artifacts (Cole *et al*., 2005; Hevner *et al*., 2004; Niehaves, 2007). Design science research as overall research approach was therefore chosen as the research paradigm for conducting the overall research.

The study considered design science research as appropriate for specifically conducting Information Visualisation research since a Visualisation project usually requires that a Visualisation artifact must be built. This proposition was communicated to the academic body of knowledge by means of two conference papers (Engelbrecht *et al*., 2014, 2015).

A number of process methods are available for conducting design science research including methods provided by Nunamaker *et al*. (1990), Purao (2002) and Peffers *et al*. (2006). From these and other possible methods the design science process model of Peffers *et al*. (2006) was chosen since it was also proposed by Hevner and Chatterjee (2010) as being appropriate for conducting design science research.

Taking the opportunities of the study, the nature of the research problem and the technical skillset of the researcher into consideration, the study suggests that the most appropriate method was chosen to answer the research problem.

- *Was the sample used in the research justifiable?*

The research study included two samples consisting of domain and IT experts. The first sample consisted of 5 experts in the domain described by the textual documents (Anglo-Boer War experts) as well as 5 IT experts. The specific software solution (a conglomeration of both the *initial* and *intermediate* software solutions) was demonstrated to the sample after which feedback was received by means of feedback questionnaires in the form of a five-point Likert scale statement set. The feedback were then analysed by using pivot tables to determine whether the software solution could be used to visually represent the coherence of facts in a textual document set. The goal was not to find the best solution, but to find a suitable solution.

The second sample consisted of the same 5 IT experts used in the software solution demonstration. The *final* model was presented to them and feedback by means of a feedback questionnaire in the form of a five-point Likert scale statement set were received. The feedback were analysed by using pivot tables to determine whether the final resulting model represented the model components required by the main research question of the study.

The decision to use 5 domain experts and 5 IT experts was based on Nielsen and Landauer's (1993) proposition that for expert user evaluations, between three and five expert users are adequate. The study concludes therefor that the chosen sample size is justifiable. Future research might include an evaluation of the solution by involving a larger number of participants.

- *Were the data analysis technique justifiable?*

The analysis of the feedback from both the specific software solution demonstrations and the *final* model presentation is discussed in Section 5.1.2 and 5.2.2. The feedback received by both samples were analysed by using pivot tables. The use of pivot tables to analyse data is a simple technique, and for the purpose of the study and the nature of the data, it is concluded that the technique is justifiable and adequate.

6.5.3 Personal reflection

Looking back at the period in my life involving the study, I can state that it was a time of many challenges and of personal growth. The challenge of balancing personal family time with study time was probably the biggest challenge. With regards to personal growth I have acquired certain academic skills which include academic writing, research and presentation skills.

At the start of my study I already possessed technical programming skills which I had acquired over a number of year. I had however very limited research experience, which was quite daunting at the time. As the study progressed I acquired a significant amount of academic research skills and experience. I am thankful to my colleagues and my employer, the CSIR, for creating the environment where I could develop and grow these skills.

During the course of the study, I particularly learned how to start off with the aim of a study and to define research questions and objectives from it. Having research questions and objectives, I learned next how to choose an appropriate research approach and methodology to achieve the required research objectives. As a result I gained a considerable amount of experience and knowledge in conducting design science research. In conducting the research I learned how to conduct a literature review to provide theoretical input to the study in order to ensure that research rigor is achieved.

In addition I did not have any academic writing or presentation skills before the study. The study produced a workshop presentation, three peer-reviewed conference papers and a peer-reviewed journal article. It was a challenge presenting the research to academic audiences over the world as I had not presented anything in public before. At the end of the research I have achieved a level of self-confidence which a quite satisfactory. As a result of presenting the research, I have also started to build a network of academic peers. I am satisfied with the value that my research study contributes to the academic community. The process of developing the final model was technically satisfying.

On a very personal, non-academic note I can add that I will always treasure the moments at home when my son used to play with his toy cars next to me while I worked on my dissertation. I will always associate the research study with my little boy.

Possible future research is discussed in the next section.

6.6 FUTURE RESEARCH

Possible future research to be conducted is discussed in this section. As a result of the feedback received from the participants during the demonstrations of the visual representation solution, possible research into providing the following future functionalities were identified:

- additional functionality to show less testimonies and facts;
- functionality to display the original untranslated source text to the user;
- functionality allowing the user to also display the coherence of facts across different source sets;
- adding functionality allowing the user to select how the facts in the dropdown list should be sorted – either alphabetically or by the number of occurrences in the document set.

In addition to future research involving additional functionality, the delineation and assumptions of the study suggests that the following future research may be conducted:

- the visual representation of the coherence of facts in a textual document set for document sets containing more than 50 documents;
- improving the accessibility and usability of the visual representation;
- investigating whether the provided software solutions is the most effective;
- an investigation into providing visual representations to assist the process of making an assumption of the truth value of facts in a textual document set based on truth theories beyond the coherence theory;
- an investigation into providing the most suitable Information Extraction method for extracting domain facts from textual documents (supervised, semi-supervised or unsupervised).

In addition future research also includes the possible application of the model to the NIKMAS system in order to visually represent the coherence of facts contained in Indigenous Knowledge descriptions. By doing so, value can be added to the research field of Indigenous Knowledge by providing a mechanism for making an assumption on the validity of selected Indigenous Knowledge descriptions.

Having provided an overview of the research process, key research findings, significance and contributions of the research, delineation and assumptions of the research, reflections on the research and possible future research, the next section concludes the dissertation.

6.7 CONCLUDING REMARKS

The study is summarised as research leading to the development of a model for the visual representation of the coherence of facts in a textual document set. The main finding was that the model could be used to assist with the process of making a decision about the possible truth value of facts in a textual document set.

The aim of the research study originated from the development of an Indigenous Knowledge capturing and protection system known as NIKMAS. In the development of the NIKMAS system, it soon became apparent that in certain Indigenous Knowledge descriptions, a coherency of facts could be identified. Furthermore, it was postulated that being able make an assumption about the coherence of facts in Indigenous Knowledge descriptions, insight into the validity of the records could be acquired. The research study therefore developed a model for a visual representation of the coherence of facts in a textual document set.

Due to intellectual property issues and the availability of Indigenous Knowledge records, an alternative textual source set was used for the development of the artifact, namely Anglo-Boer War (ABW) testimonies. The two different textual sources exhibited similar characteristics and therefore, the proposed Information Visualisation model can be applied to both Indigenous Knowledge and Anglo-Boer War (and therefore History) study fields.

In developing the model, the study's research questions and research objectives were dealt with by following a scientific research process. The result of the research process contributes to the academic computer science community, especially in the research fields of Indigenous Knowledge, Information Extraction and Information Visualisation.

Following the research study, the next step is to apply the proposed Information Visualisation model to other fields of study; in particular to the field of Indigenous Knowledge.

REFERENCES

Ackoff, R. L. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Agarwal, A. K., & Srivastava, A. (2016). Study, Comparison and Analysis of Different Types of Storage Systems. *International Journal of Current Engineering and Technology*, *6*(4), 1392 – 1399. http://doi.org/10.1145/362384.362685

Ameller, D., Collell, O., & Franch, X. (2013). The Three-Layer architectural pattern applied to plug-in-based architectures: the Eclipse case. *Software - Practice and Experience*, *43*, 391–402. http://doi.org/10.1002/spe

Auber, D., Bourqui, R., Delest, M., Lambert, A., Melan, G., Pinaud, B., … Vallet, J. (2016). *TULIP 4*.

Bihanic, D., & Polacsek, T. (2012). Models for Visualisation of Complex Information Systems. *2012 16th International Conference on Information Visualisation*, 130–135. http://doi.org/10.1109/IV.2012.32

Bischof, H.-P., Dale, E., & Peterson, T. (2006). Spiegel - A Visualization Framework for Large and Small Scale Systems. *MSV*, *6*, 199–205.

Blascheck, T., John, M., Kurzhals, K., Koch, S., & Ertl, T. (2016). VA 2 : A Visual Analytics Approach for // Evaluating Visual Analytics Applications. *IEEE Transactions on Visualizations and Computer Graphics*, *22*(1), 61–70.

Bostock, M. (2012). Circle Packing. Retrieved from http://bl.ocks.org/mbostock/4063530

Bostock, M. (2013). Data-Driven Documents. Retrieved from http://d2js.org

Bostock, M. (2013). Zoomable Circle Packing. Retrieved September 1, 2015, from http://bl.ocks.org/mbostock/7607535

Bostock, M. (2014). Gallery. Retrieved from https://github.com/mbostock/d3/wiki/Gallery

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–9. http://doi.org/10.1109/TVCG.2011.185

Burian, P. E., Rogerson, L., & Maffei III, F. R. (2010). The Research Roadmap: A Primer To The Approach And Process, *3*(8), 43–58.

Burkhard, R. A. (2004). Learning from Architects: The Difference between Knowledge Visualization and Information Visualization. In *Eight International Conference on Information Visualisation* (pp. 519–524). IEEE.

Candlish, S., & Damnjanovic, N. (1989). The Coherence Theory of Truth : Russell ' s Worst Invention ? *Ontos Verlag: Publications of the Austrian Ludwig Wittgenstein Society - New Series*, *15*.

Capuano, N., Aniello, G. D., Gaeta, A., & Miranda, S. (2015). Computers in Human Behavior A personality based adaptive approach for information systems. *Computers in Human Behavior*, *44*, 156–165. http://doi.org/10.1016/j.chb.2014.10.058

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Carlisle, J. P. (2015). Continuing the DIKW Hierarchy Conversation Continuing the DIKW Hierarchy Conversation. In *Tenth Midwest Association for Information Systems Conference*. Pittsburg, Kansas.

Carlsson, S. A. (2006). Towards an Information Systems Design Research Framework: A Critical Realist Perspective. *Design Science Research in Information Systems and Technology*, pp. 192–212. http://doi.org/10.1.1.106.7286

Castillo, D. De, & Gesmann, M. (2011). Using the Google Visualisation API with R. *The R Journal 3.2*, 40–44.

Chen, C. (2005). Visualization Viewpoints: Top 10 Unsolved Information Visualization Problems. *Computer Graphics and Applications, IEEE*, (25.4), 12–16.

Chen, M., Ebert, D., Hagen, H., Robert, S., Shneiderman, B., & Tory, M. (2009). Visualization Viewpoints. Data, Information and Knowledge in Visualization. *Computer Graphics and Applications, IEEE*, (January/February 2009), 12–19.

Chen, M., & Floridi, L. (2013). An analysis of information visualisation. *Synthese*, *190*(16), 3421–3438. http://doi.org/10.1007/s11229-012-0183-y

Chen, M., Floridi, L., & Borgo, R. (2013). What is Visualization Really for? *arXiv Preprint*, *1305.5670*(February).

Chi, E. H. (2000). A Taxonomy of Visualization Techniques using the Data State Reference Model. In *IEEE Symposium on Information Visualization* (Vol. 94301). IEEE.

Cole, R., Purao, S., Rossi, M., & Sein, M. K. (2005). Being proactive: where action research meets design research. In *ICIS 2005* (pp. 1–21).

Creswell, J. W. (2009). *Research design: Qualitative, quantitative and mixed methods approaches*. SAGE Publications. http://doi.org/10.3109/08941939.2012.723954

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Cybulska, A., & Vossen, P. (2011). Historical Event Extraction from Text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 39–43). Portland, OR, USA.

Dampier, H. (2005). Re-reading as a methodology: the case of Boer Women's testimonies. *Qualitative Research, 8(3), 367-377.*, 1899–1902.

Dang, S., & Ahmad, P. H. (2015). A Review of Text Mining Techniques Associated with Various Application Areas. *International Journal of Science and Research*, *4*(2), 2461–2466.

Devendorf, L. (2012). TopicLens: An Interactive Recommender System based on Topical and Social Connections, *i*, 41–47.

Dix, A. (2013). Introduction to Information Visualisation. In *Information Retrieval Meets Information Visualization* (pp. 1–27). Berlin, Heidelberg: Springer.

Dove, G., & Jones, S. (2012). Narrative Visualization: Sharing Insights into Complex Data. In *Interfaces and Human Computer Interaction 2012 (IHCI 2012)*. Lisbon, Portugal. Retrieved from http://openaccess.city.ac.uk/1134/

Dowden, B., & Swartz, N. (2013). Truth. In *Internet Encyclopedia of Philosophy*. Retrieved from http://www.iep.utm.edu/truth/

Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., & Jim, E. (2016). User validation in ontology alignment. In *International Semantic Web Conference* (pp. 200–217). Springer International Publishing.

Drechsler, a, & Hevner, A. (2016). A four-cycle model of IS design science research: capturing the dynamic nature of IS artifact design. *… Emerging Insights from Ongoing Design Science …*. Retrieved from https://cora.ucc.ie/handle/10468/2560

Elmqvist, N., & Fekete, J.-D. (2010). Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, *16*(3), 439–54. http://doi.org/10.1109/TVCG.2009.84

Emerson, J., Churcher, N., & Deaker, C. (2013). From Toy to Tool: Extending Tag Clouds for Software and Information Visualisation. In *22nd Australian Software*

Engineering Conference (pp. 155–164). IEEE. http://doi.org/10.1109/ASWEC.2013.27

Engelbrecht, L. (2014). Masters research in progress: Visualisation of the coherence of facts contained in a textual record set. Presentation to FASTAR/Espresso Workshop 2014. Retrieved from www.fastar.org/publications/FEW 2014 - Engelbrecht.pdf

Engelbrecht, L., & Botha, A. (2015). Designing a View for Visually Representing Information Coherence in a Document set. In *9th IDIA conference, IDIA2015, Nungwi, Zanzibar* (pp. 177–188).

Engelbrecht, L., Botha, A., & Alberts, R. (2014). Information Visualisation View Design: Principles and Guidelines. In *International Conference on Computer Science, Computer Engineering, and Social Media* (pp. 17–28). Thessaloniki, Greece.

Engelbrecht, L., Botha, A., & Alberts, R. (2015). Designing the Visualization of Information. *International Journal of Image and Graphics*, *15*(2), 1–19. http://doi.org/10.1142/S0219467815400057

Engelbrecht, L., Botha, M., & Botha, A. (2016). Using Information Visualisation to give voice to a historical community. In *13th Prato CIRN Conference*. Prato, Italy.

Eppler, M. J., & Burkhard, R. A. (2007). Visual representations in knowledge management: framework and cases. *Journal of Knowledge Management*, *11*(4), 112–122. http://doi.org/10.1108/13673270710762756

Fagin, R., Reiss, F., Kimelfeld, B., & Vansummeren, S. (2016). A Relational Framework for Information Extraction. *ACM SIGMOD Record*, *44*(4), 5–16.

Fernanda, V., Wattenberg, M., Hebert, J., Borggaard, G., Cichowlas, A., Feinberg, J., … Wren, C. R. (2013). Google+ Ripples: A Native Visualization of Information Flow. In *22nd international conference on World Wide Web* (pp. 1389–1398). International World Wide Web Conferences Steering Committee.

Figueiras, A. (2014). How to Tell Stories Using Visualization. In *18th International Conference on Information Visualisation*. IEEE.

Fogwill, T., Viviers, I., Engelbrecht, L., Krause, C., & Alberts, R. (2011). A software architecture for an indigenous knowledge management system. In *Indigenous Knowledge Technology Conference (IKTC)*. Windhoek, Namibia.

Forsell, C., & Johansson, J. (2010). An heuristic set for evaluation in information visualization. In *International Conference on Advanced Visual Interfaces - AVI '10* (p. 199). New York, New York, USA: ACM Press. http://doi.org/10.1145/1842993.1843029

Fowler, M., Rice, D., Foemmel, M., Mee, R., & Stafford, R. (2002). *Patterns of Enterprise Application*. *I Can*.

Franzosi, R., Doyle, S., McClelland, L. E., Putnam Rankin, C., & Vicari, S. (2013). Quantitative narrative analysis software options compared: PC-ACE and CAQDAS (ATLAS.ti, MAXqda, and NVivo). *Quality and Quantity*, *47*(6), 3219–3247. http://doi.org/10.1007/s11135-012-9714-3

Friese, S. (2011). FORUM : QUALITATIVE SOCIAL RESEARCH Using ATLAS . ti for Analyzing the Financial Crisis Data 1 . Introduction and Underlying Assumptions, *12*(1).

Gan, Q., Zhu, M., Li, M., Liang, T., Cao, Y., & Zhou, B. (2014). Document visualization: an overview of current research. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(1), 19–36. http://doi.org/10.1002/wics.1285

Glanzberg, M. (2013). The Concept of Truth. In *A Companion to Donald Davidson* (pp. 156 – 172). Oxford: Wiley Blackwell.

Grainger, S., Mao, F., & Buytaert, W. (2016). Environmental data visualisation for non-scientific contexts : Literature review and design framework. *Environmental Modelling and Software*, *85*, 299–318. http://doi.org/10.1016/j.envsoft.2016.09.004

Grant, C., George, C. P., Kanjilal, V., Nirkhiwale, S., Wilson, J. N., & Wang, D. Z. (2015). A Topic-Based Search, Visualization, and Exploration System. In *Twenty-Eighth International Florida Artificial Intelligence Research Society Conference* (pp. 43–48).

Gray, D. E. (2013). *Doing research in the real world*. Sage.

Gregor, S., & Hevner, A. R. (2010). Introduction to the special issue on design science. *Information Systems and E-Business Management*, *9*(1), 1–9. http://doi.org/10.1007/s10257-010-0159-8

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). TopicNets. *ACM Transactions on Intelligent Systems and Technology*, *3*(2), 1–26. http://doi.org/10.1145/2089094.2089099

Grimmer, J., & Stewart, B. M. (2012). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts ∗. *Political Analysis*, (617).

Grozev, N., & Buyya, R. (2014). Multi-Cloud Provisioning and Load Distribution for Three-Tier Applications. *ACM Transactions on Autonomous and Adaptive Systems*, *1*(June). http://doi.org/10.1145/2662112

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, *1*(1), 60–76.

Gurupur, V. P., Suh, S. C., Selvaggi, R. R., Karla, P. R., Nair, J. S., & Ajit, S. (2012). An Approach for Building a Personal Health Information System Using Conceptual Domain Knowledge. *Journal of Medical Systems*, *36*(6), 3685–3693. http://doi.org/10.1007/s10916-012-9842-2

Hall, D. (1999). *The Hall handbook of the Anglo Boer War*. (F. Pretorius & G. Torlage, Eds.). University of Natal Press.

Harger, J. R., & Crossno, P. J. (2012). Comparison of open-source visual analytics toolkits, *8294*, 82940E. http://doi.org/10.1117/12.911901

He, Y., & Grishman, R. (2015). ICE: Rapid Information Extraction Customization for NLP Novices. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 31–35). Denver, Colorado: Assocation for Computational Linguistics.

Heer, J., Card, S. K., & Landay, J. A. (2005). Prefuse : a toolkit for interactive information visualization. In *SIGCHI conference on Human factors in computing systems* (pp. 421–430). ACM.

Hermawati, S., & Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a speci fi c domain : Is there a consensus ? *Applied Ergonomics*, *56*, 34–51. http://doi.org/10.1016/j.apergo.2015.11.016

Hevner, A., & Chatterjee, S. (2010). *Design Science Research in Information Systems*. *Design Research in Information Systems* (Vol. 22). http://doi.org/10.1007/978-1-4419-5653-8

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2).

Hevner, A. R., & Chatterjee, S. (2004). Design Research in Information Systems. *Design Research in Information Systems*, *28*, 75–105. http://doi.org/10.1007/978-1-4419-5653-8

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Hofstee, E. (2009). *Constructing a Good Dissertation. A Practical Guide to Finishing a Master's, MBA or PhD on Schedule.* Johannesburg, South Africa: EPE.

Huang, D., Tory, M., Aseniero, B. A., Bartram, L., Bateman, S., Carpendale, S., … Woodbury, R. (2015). Personal Visualization and Personal Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, *21*(3), 420–433.

Hussain, M. A., Elyas, T., & Nasseef, O. A. (2013). Research paradigms: a slippery slope for fresh researchers. *Journal of Chemical Information and Modeling*, *10*(4), 2374–2381. http://doi.org/10.1017/CBO9781107415324.004

Huyser, H. W. (n.d.). Huyser Collections of Boer War testimonies, Aanwins A951. National Archives of South Africa.

Iivari, J. (2007). A Paradigmatic Analysis of Information Systems As a Design Science. *Scandinavian Journal of Information Systems*, *19*(2), 39–64. http://doi.org/10.1.1.218.2636

Jacob, N., & Thomas K., L. (1993). A mathematical model of the finding of usability problems. In *INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206–213). http://doi.org/10.1145/169059.169166

Johannesson, P., Perjons, E., & Bider, I. (2013). What Are the Siblings of Design Science. In *The 2nd International SIG Prag Workshop on IT Artefact Design & Workpractice Improvement* (pp. 1–7).

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. Pearson.

Kamins, M. A. (1993). *Secondary research: Information sources and methods* (Vol 4). Sage.

Kaushik, A., & Naithani, S. (2016). A Comprehensive Study of Text Mining Approach. *International Journal of Computer Science and Network Security*, *16*(2), 69–76.

Keim, D. A., Rossi, F., Seidl, T., Verleysen, M., & Wrobel, S. (2012). Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). *Dagstuhl Reports*, *2*(2), 58 – 83.

Keller, T., & Tergan, S. (2005). Visualizing Knowledge and Information: An Introduction. In T. Keller & S. Tergan (Eds.), *Knowledge and information visualization* (pp. 1–23). Springer Berlin Heidelberg.

Kirsch, P., Hine, A., & Maybury, T. (2015). A model for the implementation of industry-wide knowledge sharing to improve risk management practice. *SAFETY SCIENCE*, *80*, 66–76. http://doi.org/10.1016/j.ssci.2015.07.009

Krauss, S. E. (2005). Research Paradigms and Meaning Making : A Primer. *The Qualitative Report*, *10*(4), 758–770. http://doi.org/10.1176/appi.ajp.162.10.1985

Ku, C., Nguyen, J. H., & Leroy, G. (2012). TASC - Crime report visualization for investigative analysis: A case study. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)* (pp. 466–473). IEEE.

Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. *Business Management Dynamics*, *2*(10), 49–62.

Ling, X., & Weld, D. S. (2007). Temporal Information Extraction. *AAAI*, 1385–1390.

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization : recent advances and challenges. *The Visual Computer*, *30*(12), 1373–1393. http://doi.org/10.1007/s00371-013-0892-3

Lu, H., & Liu, G. (2011). Text Visualization and Visual Analytics Based on Multi-layer Topic Maps ⋆, *12*(December), 2459–2464.

Mackenzie, N., & Knipe, S. (2006). Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research*, *16*.

Mafeking. (n.d.). Retrieved March 25, 2015, from http://www2.lib.uct.ac.za/mss/bccd/Histories/Mafeking/

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251–266. Retrieved from http://www.sciencedirect.com/science/article/pii/0167923694000412

Marston, T. (2012). Design Patterns - a personal perspective. Retrieved September 3, 2015, from http://www.tonymarston.net/php-mysql/design-patterns.html

Matney, S., Avant, K., & Staggers, N. (2016). Toward an Understanding of Wisdom in Nursing. *Toward an Understanding of Wisdom in Nursing*, *21*, 1–11. http://doi.org/10.3912/OJIN.Vol21No01PPT02

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *ACM Queue*, *3.9*(November), 48–57.

McKenna, S., Mazur, D., Agutter, J., & Meyer, M. (2014). Design Activity Framework for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2191–2200.

Mooranian, M., Dillon, T., & Chang, E. (2011). Overview of cognitive visualisation. *5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011)*, 138–142. http://doi.org/10.1109/DEST.2011.5936613

Nafari, M., & Weaver, C. (2015). Query2Question : Translating Visualization Interaction into Natural Language. *IEEE Transactions on Visualization and Computer Graphics*, *21*(6), 756–769.

Nafea, S., Maglaras, L. A., Siewe, F., Smith, R., & Janicke, H. (2016). Personalized Students' Profile Based On Ontology and Rule-based Reasoning.

Niehaves, B. (2007). On Epistemological Diversity in Design Science – New Vistas for a Design-Oriented Is Research ? *Twenty-Eighth International Conference on Information Systems*, (2004), 1–13.

Nielsen, J. (1994). *Heuristic evaluation*. (J. Nielsen & R. L. Mack, Eds.). New York: John Wiley & Sons.

Nunamaker, J., Chen, M., & Purdin, T. (1990). Systems development in information systems research. *Proceedings of the Twenty-Third Annual Hawaii International Conference on System Sciences*, *7*, 89–106. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=205401

Nurnberger, A., Seising, R., & Wenzel, C. (2009). On the fuzzy interrelationships of data, information, knowledge and wisdom. In *The 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009)* (pp. 1–6). Ieee. http://doi.org/10.1109/NAFIPS.2009.5156466

O'Madadhain, J., Fisher, D., Smyth, P., White, S., & Boey, Y. (2005). Analysis and Visualization of Network Data using JUNG. *Journal of Statistical Software*, *10*(2).

Oates, B. J. (2006). *Researching Information Systems and Computing*. SAGE Publications.

Oosthuizen, J. (1998). The final nie in Afrikaans negative sentences. *Stellenbosch Papers in Linguistics (SPiL)*, *31*, 61–94. http://doi.org/10.5774/31-0-57

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The design science research process: A model for producing and presenting information systems research.

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. http://doi.org/10.2753/MIS0742-1222240302

Pretorius, F. (2011). The Boer Wars. Retrieved from http://www.bbc.co.uk/history/british/victorians/boer_wars_01.shtml

Purao, S. (2002). Design Research in the Technology of Information Systems: Truth or Dare. *Information Systems Journal*, 36. Retrieved from http://iris.nyit.edu/~kkhoo/Spring2008/Topics/DS/000DesignSc_TechISResearch-2002.pdf

Rai, P., & Murari, V. (2015). Transforming Digital Unstructured and Semi-structured Data into Structured Data with the Aid of IE and KDT. *International Journal of Research in Computer and Communication Technology*, *4*(4), 304–309.

Randjelovic, D., & Popovic, B. (2011). Visual analytics tools and theirs application in social networks analysis. In *Telecommunications Forum (TELFOR)* (pp. 1340–1343). IEEE.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. http://doi.org/10.1016/j.knosys.2015.06.015

Recupero, D. R., Nuzzolese, A. G., Consoli, S., Gangemi, A., Presutti, V., Peroni, S., & Mongiovi, M. (2015). SHELDON: Semantic holistic framework for linked ontology data. In *24th International Conference on World Wide Web* (Vol. 8982, pp. 235 – 238). ACM. http://doi.org/10.1007/978-3-319-17966-7_17

Rioth, M. J., Osterman, T. J., & Warner, J. L. (2015). Advances in Website Information Resources to Aid in Clinical Practice. *Advances in American Society of Clinical Oncology Educational book/ASCO American Society of Clinical Oncology Meeting*, *35*(2), 608–615.

Roche, W. (2013). On the Truth-Conduciveness of Coherence. *Erkenntnis*, 1–19. http://doi.org/10.1007/s10670-013-9564-y

Roller, R. ., Agirre, E. ., Soroa, A. ., & Stevenson, M. . (2015). Improving distant supervision using inference learning. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (Vol. 2, pp. 273–278). http://doi.org/10.3115/v1/P15-2045

Roller, R., & Stevenson, M. (2015). Making the most of limited training data using distant supervision. In *Workshop on Biomedical Natural Language Processing* (pp. 12–20). Beijing, China.

Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M., & Sipoš, R. (2009). Document Visualization Based on Semantic Graphs. *2009 13th International Conference Information Visualisation*, 292–297. http://doi.org/10.1109/IV.2009.57

SAICSIT 2014 Master's and Doctoral Symposium. (2014). Retrieved from http://www.saicsit2014.org

Santos, B. S., Ferreira, B. Q., & Dias, P. (2015). Heuristic evaluation in Information Visualization using three sets of heuristics : an exploratory study. In *International Conference on Human-Computer Interaction* (pp. 259–270). Springer International Publishing.

Satyanarayan, A., & Heer, J. (2014). Lyra : An Interactive Visualization Design Environment. In *Eurographics Conference on Visualization (EuroVis)* (Vol. 33, pp. 351–360). http://doi.org/10.1111/cgf.12391

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012). Open language learning for information extraction. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 523–534). Assocation for Computational Linguistics.

Schreiber, M. ., Kraft, B. ., & Zündorf, A. . (2016). Cost-efficient quality assurance of natural language processing tools through continuous monitoring with continuous integration. In *3rd International Workshop on Software Engineering Research and Industrial Practice, SER and IP 2016* (pp. 46–52). http://doi.org/10.1145/2897022.2897029

Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English Language Teaching*, *5*(9), 9–16. http://doi.org/10.5539/elt.v5n9p9

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. http://doi.org/10.1109/VL.1996.545307

Smit, B. (2014). Teaching Atlas . ti ™ in South Africa : Reflections on Doctoral Preparedness, *40*(2), 213–217.

Strötgen, J., Gertz, M., & Popov, P. (2010). Extraction and exploration of spatio-temporal information in documents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10* (p. 1). New York, New York, USA: ACM Press. http://doi.org/10.1145/1722080.1722101

The Boer War. (2008). Retrieved November 12, 2016, from http://www.australia.gov.au/about-australia/australian-story/boer-war

Tremblay, M. C., Parra, C., & Castellanos, A. (2015). Analyzing Corporate Social Responsibility Reports Using Unsupervised and Supervised Text Data Mining. In *International Conference on Design Science Research in Information Systems* (pp. 439 – 446). Springer International Publishing.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. http://doi.org/10.1016/j.ipm.2013.08.006

Van Biljon, J., & Renaud, K. (2015). Facilitating Knowledge Visualisation as Communication and Knowledge Transfer Mechanism in Postgraduate Learning. In *International Conference on Mobile and Contextual Learning* (Vol. 1, pp. 156–

171). Springer International Publishing. http://doi.org/10.1007/978-3-319-25684-9

Van Heyningen, E. (2009). The Concentration Camps of the South African (Anglo-Boer) War, 1900-1902. *History Compass*, *7*(1), 22–43. http://doi.org/10.1111/j.1478-0542.2008.00562.x

Verma, V. K., Ranjan, M., & Mishra, P. (2015). Text Mining and Information Professionals Role, issues and Challenges. In *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on* (pp. 133–137). IEEE. http://doi.org/10.1109/ETTLIS.2015.7048186

Vijayarani, S., Ilamathi, J., & Nithya. (2015). Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16. Retrieved from http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf

Vosloo, I., & Kourie, D. G. (2008). Server-centric Web frameworks. *ACM Computing Surveys*, *40*(2), 1–33. http://doi.org/10.1145/1348246.1348247

Wahyuni, D. (2010). The research design maze: understanding paradigms,cases,methods and methodologies. *Journal of Applied Manage Ment Accounting Research*, *10*(1), 69–80.

Walny, J., Huron, S., & Carpendale, S. (2015). An Exploratory Study of Data Sketching for Visual Representation. In *Eurographics Conference on Visualization (EuroVis)* (Vol. 34, pp. 231–240).

Wang, R., Perez-Riverol, Y., Hermjakob, H., & Vizca, J. A. (2015). Open source libraries and frameworks for biological data visualisation : A guide for developers. *Proteomics*, *15*(8), 1356–1374. http://doi.org/10.1002/pmic.201400377

Wang, Y. (2015). Formal Cognitive Models of Data, Information, Knowledge, and Intelligence. *WSEAS TRANSACTIONS on COMPUTERS*, *14*, 770–81. Retrieved from http://www.ucalgary.ca/icic/

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text. *ACM Computing Surveys*, *44*(4), 1–36. http://doi.org/10.1145/2333112.2333115

Wu, L., & Hsu, P. (2013). An interactive and flexible information visualization method. *Information Sciences*, *221*, 306–315. http://doi.org/10.1016/j.ins.2012.09.038

Wu, Y., Lehman, A. R., & Dunaway, D. J. (2015). Evaluations of a Large Topic Map as a Knowledge Organization Tool for Supporting Self-Regulated. *Knowledge Organization*, *42*(6), 386–398.

Wu, Y., Liu, X., Liu, S., & Ma, K. (2013). ViSizer : A Visualization Resizing Framework. *IEEE Transactions on Visualization and Computer Graphics*, *19*(2), 278–290.

Yang, T., Torget, A., & Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 96–104). Portland, OR, USA.

Yeung, C. A., & Jatowt, A. (2011). Studying How the Past is Remembered: Towards Computational History through Large Scale Text Mining, 1231–1240.

Young, J. O. (2013). The Coherence Theory of Truth. In *The Stanford Encyclopedia of Philosophy* (Summer 201). Retrieved from http://plato.stanford.edu/archives/sum2013/entries/truth-coherence/

Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *2015-Novem*, 681–685. http://doi.org/10.1109/ICSESS.2015.7339149

Zhao, H., & Lu, L. (2015). Variational Circular Treemaps for Interactive Visualization of Hierarchical Data. In *IEEE Pacific Visualization Symposium (PacificVis)* (pp. 81–85). IEEE.

Zuk, T., Schleiser, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). Heuristics for information visualization evaluation. In *AVI Workshop on Novel Evaluation Methods for Information Visualization* (pp. 1–6).

APPENDIX A: PARTICIPANT CONSENT FORM

CONSENT FORM

A model for a visual representation of unstructured, textual data towards visually representing the coherence of facts contained in a data set

Dear Mr/Mrs/Miss/Ms _____ Date...../...../20...

NATURE AND PURPOSE OF THE STUDY

A large amount of data is contained in textual records. This data is unstructured and to visualise its content is not trivial. In order to visualise data contained in unstructured textual records, it must first be transformed into a structured format. This research aims to visualise the coherence of facts contained in textual sources in order to make an assumption on the validity of the textual records as a set. The output of this research is a model that abstracts the process into components required to transform unstructured, textual records into a structured format that can be used to visually represent the coherence of facts contained in the textual record set. The model provides guidance and practices for performing similar tasks on textual record sets containing secondary data. This research is conducted by following the design science research methodology.

RESEARCH PROCESS

1. The research study requires that a developed Information System (IS) artifact be demonstrated to a small number Information Technology (IT) professionals and/or domain experts in the field of South African history.
2. As part of the demonstration, participants will be required to complete a short questionnaire.
3. The completed questionnaires will be used as input into a process of making possible adjustments to the artifact in order to improve it and determine whether it is perceived as being useful.

CONFIDENTIALITY

The opinions of the participants are viewed as strictly confidential, and only members of the research team will have access to the information. No data published in dissertations and journals will contain any information through which participants may be identified. Your anonymity is therefore ensured.

WITHDRAWAL CLAUSE

I understand that I may withdraw from the study at any time. I therefore participate voluntarily until such time as I request otherwise.

POTENTIAL BENEFITS OF THE STUDY

The research study has several significant benefits. In conducting this research a better understanding of textual content analysis, information extraction and data visualisation techniques and options will be acquired. The study will demonstrates how visualisation can be used to make an assumption on the validity of a data set. This assumption based on the coherence between domain facts contained in a set of textual records, indicates that some truth is probably contained in the record set.

The problem of unstructured textual data and the visualisation of it is not just relevant to historical studies but it is also relevant in other fields of studies, for instance in indigenous knowledge studies (traditional medicine), medical studies (health records) and political studies (political texts). The study can therefore be the starting point for similar research in fields other than history.

INFORMATION

If I have any questions concerning the study, I may contact the supervisor, Dr. Adele Botha, at the Council for Scientific and Industrial Research (CSIR), Pretoria, Tel: 012 841 3265.

CONSENT

I, the undersigned, …… (full name) have read the above information relating to the project and have also heard the verbal version, and declare that I understand it. I have been afforded the opportunity to discuss relevant aspects of the project with the project leader, and hereby declare that I agree voluntarily to participate in the project.

I indemnify the university and any employee or student of the university against any liability that I may incur during the course of the project.

I further undertake to make no claim against the university in respect of damages to my person or reputation that may be incurred as a result of the project/trial or through the fault of other participants, unless resulting from negligence on the part of the university, its employees or students.

I have received a signed copy of this consent form.

Signature of participant: .....

Signed at …… on ……

WITNESSES

1 .....

2 .....

APPENDIX B: ETHICAL CLEARANCE CERTIFICATE

UNISA | college of science, engineering and technology

Dear Mr Louis Engelbrecht (31913482)

Date: 2015-01-27

Application number:
012/VRP/2015

REQUEST FOR ETHICAL CLEARANCE: (A model for visual representation of unstructured, textual data towards visually representing the coherence of facts contained in a data set)

The College of Science, Engineering and Technology's (CSET) Research and Ethics Committee has considered the relevant parts of the studies relating to the abovementioned research project and research methodology and is pleased to inform you that ethical clearance is granted for your research study as set out in your proposal and application for ethical clearance.

Therefore, involved parties may also consider ethics approval as granted. However, the permission granted must not be misconstrued as constituting an instruction from the CSET Executive or the CSET CREC that sampled interviewees (if applicable) are compelled to take part in the research project. All interviewees retain their individual right to decide whether to participate or not.

We trust that the research will be undertaken in a manner that is respectful of the rights and integrity of those who volunteer to participate, as stipulated in the UNISA Research Ethics policy. The policy can be found at the following URL:
http://cm.unisa.ac.za/contents/departments/res_policies/docs/ResearchEthicsPolicy_apprvCounc_21Sept07.pdf

Please note that the ethical clearance is granted for the duration of this project and if you subsequently do a follow up study that requires the use of a different research instrument, you will have to submit an addendum to this application, explaining the purpose of the follow-up study and attach the new instrument along with a comprehensive information document and consent form.

Yours sincerely

Prof Ernest Mnkandla
Chair: College of Science, Engineering and Technology Ethics Sub-Committee

Prof IOG Mocha
Executive Dean: College of Science, Engineering and Technology

RECEIVED
2015 -01- 30
Office of the Executive Dean
College of Science, Engineering and Technology

University of South Africa
College of Science, Engineering and Technology
The Science Campus
C/o Christiaan de Wet Road and Pioneer Avenue,
Florida Park, Roodepoort
Private Bag X6, Florida, 1710
www.unisa.ac.za/cset

UNISA | college of science, engineering and technology

APPENDIX C: BACKGROUND INFORMATION ON THE HUYSER COLLECTION AND THE ANGLO-BOER WAR

The following is a short summary of the Anglo-Boer War as published on the official website of the Australian website ("The Boer War", 2008):

> "The Boer War took place in the area that is now South Africa. The British held a territory they called Cape Colony. The Boers, descendants of Dutch settlers, held Transvaal and Orange Free State.
>
> The main commercial and strategic reasons for the war were that the British wanted the gold mines of the Boer states. A British raid in 1896 on the Boers, called the Jameson Raid, resulted in the Boers declaring war on Britain on 11 October 1899. Despite capturing a number of British towns early on in the campaign, the Boers lost their strategic offence early in the conflict. Subsequently, the Boer commandos turned to guerrilla warfare, which the British responded to in kind, and the war became a brutal and drawn-out affair for the next two years.
>
> The British reinforced their army and received support from Australia, Canada, New Zealand, Ceylon (now Sri Lanka) and India. Many battles and sieges were fought. Under the command of Lord Kitchener, the British gained control in open battle by late 1900.
>
> Eventually Britain won the war and, with the signing of the Treaty of Vereeniging on 31 May 1902, all Boers became British subjects. However, within nine years South Africa became a self-governing dominion led by former Boer generals."

During the guerrilla warfare phase of the war, the British forces responded by following a scorched earth policy whereby some towns and thousands of farmsteads (including livestock and crops) were destroyed. The women and children whose homes were destroyed were then confined into concentration camps (Pretorius, 2011). By February 1902 the concentration camps housed about 215000 civilians of which about 42000 died in just a few months (Hall, 1999).

About 27 years after the war, H.W. Huyser collected a set of 37 testimonies (in the form of informal narratives) from survivors of the concentration camp which was located near the present day Mahikeng (previously known as Mafikeng). These testimonies are currently housed at the South African National Archives in Pretoria (Huyser, n.d.). Their reliability have been questioned by various authors including Van Heyningen (2009) and Dampier (2005).