DOCUMENT AND QUERY EXPANSION METHOD WITH DIRICHLET SMOOTHING MODEL FOR RETRIEVAL OF METADATA CONTENT IN DIGITAL RESOURCE OBJECTS

WAFA’ ZA’AL MOHAMMAD ALMA’AITAH

UNIVERSITI SAINS MALAYSIA

2020

DOCUMENT AND QUERY EXPANSION METHOD WITH DIRICHLET SMOOTHING MODEL FOR RETRIEVAL OF METADATA CONTENT IN DIGITAL RESOURCE OBJECTS

by

WAFA' ZA'AL MOHAMMAD ALMA'AITAH

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

March 2020

ACKNOWLEDGEMENT

First and foremost, I am forever thankful to Almighty Allah, the most gracious and the most merciful because if not for his blessing I would have never had this opportunity and capability to proceed successfully to complete my PhD thesis. This thesis would not have been possible without the inspiration, unconditioned love and support from my parents, Mr. Za'al Al ma'aitah and Mrs. Aisha Al ma'aitah. My thanks and appreciation to both of them for being part of this journey and making this thesis possible. I am forever indebted to them for giving me the opportunities and experiences that have made me who I am.

My most heartfelt gratitude goes to my supervisor, Professor Dr. Abdullah Zawawi Talib, for his kind and constant help, wise guidance, stimulating feedback and encouragement during these past four years. He is someone you will instantly love and never forget once you meet him. He has provided insightful discussions on the research, scientific advice, knowledge and suggestions. He is my primary resource for getting my scientific questions answered. I hope that I could be as lively enthusiastic and energetic as he is. My thanks also go to my co-supervisor Mr. Mohd. Azam Osman.

I would like to acknowledge that my research is partially supported by Universiti Sains Malaysia under the Research University Grant -Account no. 1001/PKOMP/8014075. Last but not least, I would like to thank my sisters, brothers, nieces and nephews for supporting and sharing a wonderful time throughout writing this thesis and with my life in general

TABLE OF CONTENTS

ACKNOWLEDGEMENT………………………………………………………. ii

TABLE OF CONTENTS ……………………………………………….. iii

LIST OF TABLES ……………………………….…………………… ix

LIST OF FIGURES…………………………………………................ xii

LIST OF ABBREVIATIONS…………………………………………… xiv

ABSTRAK ………………………………………………………………….. xvi

ABSTRACT ………………………………………………………………… xviii

CHAPTER 1 - INTRODUCTION

1.1 Metadata ……………………………………………………………… 1

1.2 Digital Resource Objects …………………………………………….. 2

1.3 Digital Cultural Heritage Collections …………………………....... 3

1.4 Information Retrieval ………………………………………………… 6

1.4.1 Information Retrieval Expansion Methods …………………... 6

1.4.1(a) Document Expansion Method……………………... 7

1.4.1(b) Query Expansion Method……………………….…. 7

1.4.2 Information Retrieval Models ……………………………….... 8

1.4.2(a) Language Model …………………………………... 9

1.5 Motivation …………………………………………………………… 10

1.6 Problem Statement ………………………………………………….. 10

1.7 Research Objectives ………………………………........................ 12

1.8 Research Contributions ……………………………………………… 13

1.9 Scope and Limitations ………………………………………………….….. 13

1.10 Thesis Organization …………………………………………............... 14

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction ……………..…………………………………………… 16

2.2 Data Retrieval …………………………………………………….... 18

2.2.1 Mapping Generic Schema…………………………………….. 20

2.2.2 Mappings across Existing Schemas ………………………... 21

2.2.3 Ontology Mapping Schemas ………………………………..... 22

2.3 Information Retrieval ……………………………………………… 25

2.3.1 Document Expansion Method …………………..…………..... 25

2.3.1(a) Co-occurrence Resources Approach ……………… 26

2.3.1(b) External Resources Approach ……………………… 27

2.3.1(c) DE Methods in DROs ………………………………. 28

2.3.2 Query Expansion Method ………………………………….... 30

2.3.2(a) Local Analysis ……………………………………...... 31

2.3.2(b) Global Analysis …………………………………..... 32

2.3.2(c) Semantic Analysis ………………………………….. 32

2.3.2(d) QE Methods in DROs ……………………………… 34

2.3.3 Retrieval Model ……………………………………………… 38

2.3.3(a) N-gram Models …………………………………... 39

2.3.3(a)(i) Unigram Model. ……………................ 40

2.3.3(b)(ii) Bigram Model ……………………….….. 40

2.3.3(b) Smoothing Models ……………………………..... 42

2.3.3(b)(i) Jelinek-Mercer Smoothing Model…... 42

2.3.3(b)(ii) Dirichlet Smoothing Model …………. 43

2.3.3(b)(iii) Absolute Discounting Smoothing Model 44

2.3.3(c) Estimation Models ………………………………... 45

2.3.3(c)(i) Query Likelihood Retrieval Model …… 45

2.3.3(c)(ii) Document Likelihood Retrieval Model 47

2.3.3(c)(iii) Comparative Model ………………… 47

2.3.3(c)(iv) Cluster-based Language Model ……….. 48

2.3.3(c)(v) Further Studies on LM …………….…. 49

2.4 Data Retrieval versus Information Retrieval in DROs …………….…. 52

2.5 Discussion ……………………………………………………………. 54

2.6 Summary …………………………………………………………... 57

CHAPTER 3 - RESEARCH METHODOLOGY

3.1 Introduction ……………………………………………………….... 59

3.2 Overview of the Methodology ……………………………………... 59

3.3 Overview of Dataset Benchmark Collection …………………………. 62

3.4 Using Wikipedia ……………………………………………………..... 65

3.5 Performance Evaluation Measurements ……………………………… 65

3.5.1 Statistical Test ……………………………………………... 67

3.6 Summary …………………………………………………………... 68

CHAPTER 4 - ENHANCED DOCUMENT EXPANSION METHOD

4.1 Introduction ……………………………………………………….... 70

4.2 Traditional Document Expansion Method ……………………………. 71

4.3 Proposed Enhanced Document Expansion Method ………………….. 73

4.3.1 Retrieval ……………………………………………………….... 75

4.3.2 Division………………………………………………….…...... 79

4.3.3 Estimation ……………………………………….…..…..…. 80

4.3.4 Integration …………………………………………………….. 82

4.4 Experiments, Results and Discussion ………………………………… 83

4.4.1 Experiment Setup ……………………………………............. 84

4.4.2 Results and Discussion …………………………………….…... 88

4.4.2(a) E_C Method ………………………………………. 89

4.4.2(b) DE_S Method …………………………….………….. 93

4.4.2(c) DE_B Method ………………………….……………. 95

4.4.2(d) Comparison among the Proposed Methods ……… 96

4.5 Statistical Significance Analysis ……………………………………… 97

4.6 Summary ……………………………………………………..…… 98

CHAPTER 5 - ADAPTIVE STRUCTURED DIRICHLET SMOOTHING MODEL

5.1 Introduction ……………………………………………………..… 100

5.2 Fundamental of Dirichlet Smoothing Model ……………………….... 101

5.3 Adaptive Dirichlet Smoothing Model ……………………..………. 105

5.4 Structured Dirichlet Smoothing Model ……………………………….. 108

5.5 Adaptive Structured Dirichlet Smoothing Model ……………..……… 111

5.6 Experiments, and Results and Discussion ……………………..…. 113

5.6.1 Experiment Setup …………………………………..… 113

5.6.2 Result based on the ADS Model ……………...……… 114

5.6.3 Result based on the SDS Model …………………….... 116

5.6.4 Result based on the ASDS Model ……………………. 117

5.6.5 Comparison among the Proposed Models …………..... 119

5.6.6 Predicted Probability of Relevance ……………….…... 121

5.7 Statistical Significance Analysis ……………………………………… 124

5.8 Summary ……….………………………………………………...….... 125

CHAPTER 6 - SEMANTIC QUERY EXPANSION METHOD

6.1 Introduction ……………………………………………………...…… 127

6.2 Overview of Query Expansion ……………………………………..… 129

6.3 Proposed Semantic Query Expansion Method …………………..….. 131

6.3.1 Query Terms Definition ……………………………………… 134

6.3.2 Identifying the Candidate Terms ……………………….…….. 135

6.3.3 Correlation Algorithm…………………………………….…... 136

6.4 Experiments, Results and Discussion…………………………………. 139

6.4.1 Experiment Setup…………………………………………….... 140

6.4.2 Experiment Results………………………………….......... 142

6.4.3 Statistical Significance Analysis…………………………..…. 144

6.5 Summary ……………………………………………………………... 145

CHAPTER 7 - INFORMATON RETRIEVAL FRAMEWORK FOR DIGITAL RESOURCE OBJECTS

7.1 Introduction ……………………………………………………………... 147

7.2 Overview of the Proposed IR Framework …………………………….. 148

7.3 Proposed IR Framework …………………………………..….…..……… 148

7.3.1 ASDS Model and IRF Components ………………………….. 151

7.4 Experiments, Results and Discussion……………………………….……… 152

7.4.1 Experiment Setup……………………………………………..…. 152

7.4.2 Experiment Results…………………………………………… 153

7.4.3 Statistical Significance Analysis…………………………...…. 159

7.5 Summary ………………………………………………………... 163

CHAPTER 8 - CONCLUSION AND FUTURE WORK

8.1 Conclusion………………………………………………….…..... 164

8.2 Future Work ………………………….………………………… 166

REFERENCES ………………………………….……………..……… 168

APPENDICES

APPENDIX A - REAL EXAMPLE OF ADAPTIVE STRUCTURED DIRCHLIT MODEL

APPENDIX B - REAL METADATA UNITS

APPENDIX C - TEST QUERIES

LIST OF PUBLICATIONS

LIST OF TABLES

		Page
Table 2.1	Limitations of DR mapping approaches in DROs	23
Table 2.2	Summary of DR in DROs	24
Table 2.3	Summary of document expansion methods	29
Table 2.4	Summary of comparison between global, local and semantic analyses	34
Table 2.5	Summary of query expansion method	36
Table 2.7	Comparison between unigram and bigram models	42
Table 2.8	Comparison of smoothing models	45
Table 2.9	Summary of advantages and disadvantages of the estimation models	49
Table 2.10	Summary of Comparison between IR and DR	53
Table 3.1	CHiC collections by language and media type (Petras, 2013)	63
Table 4.1	Comparison of MAP performance for retrieving CHiC2013 based on: vector space model, probabilistic model and language model	85
Table4.2	Language model setting	86
Table 4.3	Statistics of the test collection	86
Table 4.4	Summary of DE methods	87
Table 4.5	MAP and P@10 for CHiC2013, ECHiC2013_DE, and ECHiC2013_EDE.	88
Table 4.6	MAP and P@10 for CHiC2013, ECHiC2013_DE, ECHiC2013_DE_C1 and ECHiC2013_DE_C2	50
Table 4.7	Terms weight on both Eq. (4.1) and Eq. (4.3)	92
Table 4.8	MAP and P@10 for CHiC2013, ECHiC2013_DE, and ECHiC2013_DE_S	94

Table 4.9 MAP and P@10 for CHiC2013, ECHiC2013_DE, and ECHiC2013_DE_B model 95

Table 4.10 Overall performance of the proposed methods 97

Table 4.11 Paired t-test analysis on both measures MAP and P@10 for (ECHiC2013_EDE, ECHiC2013_DE_S, ECHiC2013_DE_C and ECHiC2013_DE_B) against the benchmarks (ECHiC2013_DE and ECHiC2013) 98

Table 5.1 MAP and P@10 for DS, EDS+ and DS model 115

Table 5.2 MAP and P@10 for DS, EDS+, and SDS model 116

Table 5.3 MAP and P@10 for DS, EDS+, and ASDS model 118

Table 5.4 μ values for some sample queries over 10 documents 119

Table 5.5 Overall performance of the proposed models 120

Table 5.6 The probability values for relevance and non-relevance results related to "silent film" query based on DS and ADS model 122

Table 5.7 The probability values for relevance and non-relevance results related to "silent film" query based on SDS and ASDS model 123

Table 5.8 Paired t-test analysis on both measures MAP and P@10 for ADS, EDS, and ASDS 125

Table 6.1 Examples of user's queries and their expansions 141

Table 6.2 Statistics of the test queries 142

Table 6.3 MAP and P@10 for IRS_1, IRS_2, IRS_3, and IRS_4 143

Table 6.4 Paired t-test analysis on both measures MAP and P@10 for IRS_1, IRS_2, IRS_3, and IRS_4 144

Table 7.1 Summary of the proposed IRF versions and their components 150

Table 7.2 MAP and P@10 for the proposed IRF versions over the first benchmarks group (DE, DS and QE) 154

Table 7.3 MAP and P@10 for the proposed IRF versions over the second benchmark group (EDE, ASDS and SQE) 155

Table 7.4 Paired t-test analysis on both measures (MAP and P@10) for the proposed IRF versions over the first benchmarks group (DE, DS and QE) 161

Table 7.5 Paired t-test analysis on both measures (MAP and P@10) for the proposed IRF versions over the second benchmarks group (EDE, ASDS and SQE) 162

LIST OF FIGURES

		Page
Figure 2.1	An overview of the literature review	17
Figure 3.1	Research methodology	60
Figure 4.1	Enhanced document expansion (EDE) method	74
Figure 4.2	Sample enriched metadata content unit by the EDE method	83
Figure 4.3	pre-processing step	84
Figure 4.4	Comparison of CHiC2013, ECHiC2013_DE, and ECHiC2013_EDE using averaged 9-point Precision-Recall curve	88
Figure 4.5	Comparison of CHiC2013, ECHiC2013_DE, ECHiC2013_DE_C1 and ECHiC2013_DE_C2 using averaged 9-point Precision-Recall curve	91
Figure 4.6	Comparison of CHiC2013, ECHiC2013_DE, and ECHiC2013_DE_S using averaged 9-point Precision-Recall curve	95
Figure 4.7	Comparison of CHiC2013, ECHiC2013_DE, and ECHiC2013_DE_B using averaged 9-point Precision-Recall curve	96
Figure 5.1	Methodology of the experiments	113
Figure 5.2	Comparison of DS and EDS+ with ADS using averaged 9-point Precision-Recall curve	115
Figure 5.3	Comparison of SDS with DS and EDS+ using averaged 9-point Precision-Recall curve	117
Figure 5.4	Comparison of ASDS with DS and EDS+ using averaged 9-point Precision-Recall curve	118
Figure 5.5	Comparison among the proposed models using averaged 9-point Precision-Recall curve	120
Figure 6.1	Semantic query expansion method	132
Figure 6.2	Methodology of the experiment (the labels refer to the branch number)	140

Figure 6.3	Comparison of IRS _1, IRS SI_2, IRS _3 and IRS_4 using averaged 9-point Precision-Recall curve	143
Figure 7.1	Proposed IR framework for DRO documents	149
Figure 7.2	Comparison of the IRF versions ((a)-(h)) with the first benchmark group (DE, DS, and QE) using averaged 9-point Precision-Recall curve	156
Figure 7.3	Comparison of IRF versions ((a)-(h)) with the second benchmark group (EDE, ASDS, and SQE) using averaged 9-point Precision-Recall curve	158

LIST OF ABBREVIATIONS

ADS	Adaptive Dirichlet Smoothing
ASDS	Adaptive Structured Dirichlet Smoothing
CH	Cultural Heritage
CRM	Conceptual Reference Model
DE	Document Expansion
DE_B	Document Expansion With Backup
DE_C	Document Expansion With Content
DE_S	Document Expansion With Sentences
DP	Dirichlet Prior
DR	Data Retrieval
DROs	Digital Resource Objects
DS	Dirichlet Smoothing
EDE	Enhanced Document Expansion
EDM	Europeana Data Model
GA	Global Analysis
IMDF	Inverse Metadata Frequency
IR	Information Retrieval
IRF	IR Framework
IRF_V	IRF Versions
LA	Local Analysis
LM	Language Model
MAP	Mean Average Precision
NLP	Natural Language Processing

P@10	Precision At Ten Documents
QE	Query Expansion
QLEM	Query Likelihood Estimation Model
SDS	Structured Dirichlet Smoothing
SQE	Semantic Query Expansion
TF-IDF	Term Frequency-Inverse Document Frequency

KAEDAH PENGEMBANGAN DOKUMEN DAN PERTANYAAN DENGAN MODEL PELICINAN DIRICHLET UNTUK DAPATAN SEMULA KANDUNGAN METADATA DALAM OBJEK SUMBER DIGITAL

ABSTRAK

Objek sumber digital (DRO) merujuk kepada maklumat yang berstruktur, yang menghuraikan, menerangkan, dan memudahkan dapatan semula, penggunaan dan pengurusan sumber maklumat. Kebelakangan ini, keperluan untuk mencapai kandungan DRO telah ditangani secara berbeza oleh komuniti penyelidikan dapatan semula data (DR) dan dapatan semula maklumat (IR). DR didapati tidak memadai dalam menyediakan kandungan metadata yang diperkaya dan mungkin gagal untuk meningkatkan prestasi dapatan semula. Tesis ini mencadangkan kerangka IR yang terdiri daripada tiga tahap utama: kaedah pengembangan dokumen dipertingkat (EDE), model pelicinan Dirichlet berstruktur boleh suai (ASDS) dan kaedah pengembangan pertanyaan semantik (SQE). Tahap pertama melibatkan kaedah EDE yang mana sebuah tatacara baru diperkenalkan untuk meningkatkan kandungan maklumat mengikut beberapa langkah tertentu dengan menambah maklumat baru yang lebih relevan dan lebih dekat kepada setiap unit metadata dalam setiap dokumen manakala tahap kedua melibatkan model ASDS yang mempunyai dua senario untuk menambah baik model pelicinan Dirichlet. Senario pertama adalah untuk meningkatkan model dengan mengambil kira struktur dokumen seperti dalam pelicinan Dirichlet berstruktur (SDS) yang dicadangkan manakala senario kedua adalah untuk mengubah parameter yang digunakan dalam model seperti dalam model pelicinan Dirichlet boleh suai (ADS) yang dicadangkan. Tahap ketiga kerangka yang dicadangkan ini melibatkan kaedah SQE yang dicadangkan untuk meningkatkan prestasi dapatan semula DRO dengan menambah baik mutu kata-kata calon yang ditambah secara semantik kepada keseluruhan kata pertanyaan. Eksperimen yang ekstensif dilakukan untuk menilai keberkesanan kaedah-kaedah, model dan kerangka IR yang dicadangkan menggunakan koleksi CHiC2013 yang tersedia kepada khalayak ramai. Keputusan eksperimen menunjukkan bahawa prestasi kaedah EDE, model ASDS, kaedah SQE dan kerangka IR yang dicadangkan masing-masing bertambah baik sebanyak 10.5%, 11.3%, 8.1%, dan 25.7% (ukuran kejituan purata min) ke atas kaedah-kaedah, model-model dan kerangka-kerangka konvensional.

DOCUMENT AND QUERY EXPANSION METHOD WITH DIRICHLET SMOOTHING MODEL FOR RETRIEVAL OF METADATA CONTENT IN DIGITAL RESOURCE OBJECTS

ABSTRACT

Digital resource objects (DRO) refer to information that are structured which elaborate, describe, and ease retrieval, usage and management of information resources. Lately, the need for accessing the content of DROs has been addressed differently by data retrieval (DR) and information retrieval (IR) research communities. DR is found to be inadequate in providing enriched metadata content and may fail to enhance the retrieval performance. In this thesis, an IR framework is proposed which consists of three main stages: enhanced document expansion (EDE) method, adaptive structured Dirichlet smoothing (ASDS) model, and semantic query expansion (SQE) method. The first stage involves proposing an EDE method in which a new procedure is introduced to increase each metadata unit content according to some specific steps by adding new information which is more relevant and closer to each metadata unit in each document while the second stage involves proposing an ASDS model that has two scenarios to improve the Dirichlet smoothing model. The first scenario is to enhance the model by taking into account of the document structure as in the proposed structured Dirichlet smoothing (SDS) model while the second scenario is to modify the parameters used in the model as in the proposed Adaptive Dirichlet smoothing (ADS) model. The third stage of the proposed framework involves the proposed SQE method to enhance the retrieval performance of DROs by improving the quality of candidate terms that are added semantically to the entire query term. Extensive experiments were conducted to evaluate the effectiveness of the proposed methods, model and IR framework using the publicly available CHiC2013 collection. The experimental results show that the performances of the proposed EDE method, ASDS model, SQE method and IR framework improve by 10.5%, 11.3%, 8.1%, and 25.7% (mean average precision measure) respectively over conventional methods, models and frameworks.

CHAPTER 1

INTRODUCTION

1.1 Metadata

Metadata refer to data that speak in volumes about data, and serve as a tool that aids users, seekers and owners with information resources. These characteristics amplify the significance of finding and managing metadata. Metadata are data that describe the format, attributes and content of an information resource or data record. In fact, metadata have been applied to describe highly structured information resources such as text documents (Barker, 2005). Metadata can be categorised into three groups (Riley, 2017): (1) Descriptive metadata: describe resources for identification and discovery purposes that are inclusive of certain elements, for instance, keywords, title, authors, and abstract; (2) Structural metadata: depict the ways compound objects can be placed together, such as the ways of arranging pages to create chapters; (3) Administrative metadata: offer information to manage resources, for example details regarding technical information, accessibility as well as file creation and type.

The metadata schemes which are known as schema refer to a set of metadata features meant to serve a particular function, for example, detailing certain source of information. The metadata features are defined as semantics of scheme. Hence, the schemes of metadata specify the features along with their semantics. In fact, metadata schemes can determine formulation of content rules (e.g. determination of the main title), representation rules for the content (e.g. rules for capitalisation) as well as permissible content values (e.g. terms for specific vocabulary) (Coyle, 2005).

Recently, the need for accessing metadata content has been addressed differently via data retrieval (databases) (DR) and information retrieval (IR) research communities. IR assists a user in accessing the desired information (Patel *et al.*, 2005). Nevertheless, it is not easy to formalise user's information as the information has to be transformed into a query that can be processed by the IR system. Relevant information can be retrieved through the IR system by using the user's query.

Meanwhile, the DR systems identify objects that meet the defined conditions as those in a relational algebra expression (Maurya *et al.*, 2013). Total failure in a database system refers to inaccurate retrieval of any object. Hence, IR addresses issues related to content interpretation in determining relevancy of the information while DR looks into performance issues, data models that are well-defined, and languages of expressive query. It is notable that in accessing and retrieving metadata content with special structure, it is essential to make the content easily available and more accessible to the users. Besides, there is a need for effective IR systems to handle descriptive metadata content retrieval.

1.2 Digital Resource Objects

Digital resource object **(**DRO) refers to information that is structured which elaborates, describes and eases the retrieval, usage and management of information resources (Witten *et al.*, 2002). Apart from the content storage, DROs offer platforms to seek, retrieve and organise contents from databases. Standardised description of resources aids in the discovery and retrieval of information resources in digital format by describing individual files, single objects or complete collections (Kalisdha & Suresh, 2017). DROs use metadata as data schemas or elements to describe the contents. Hence, metadata can be considered as digital objects for application at varying levels of aggregation that may be incorporated into DROs or separately stored. According to Aruleba *et al.* (2016), metadata should: (1) conform to community-based standards, (2) be appropriate to the materials being described and the needs of users, and (3) support interoperability and facilities to expand access. Thus, an essential part of the digital resources is as digital cultural heritage collections. As a result, many cultural heritage organisations such as galleries, libraries, archives and museums have moved towards massive digitisation of information to secure long-term preservation of valuable archived materials (Pattuelli, 2011).

Unlike traditional objects held in organisation's records, DRO's content can be shared, combined, and aggregated online and the content of digital files can be easily modified as well. These features provide a number of benefits for users of digitised content in enhancing access to digital collections and allowing their reuse for research purposes, learning and developing new commercial content. However, the ease of sharing and introducing changes to digital files complicates the task of protecting personal information contained in digitized documents and ensuring their trustworthiness (Er *et al.*, 2018; Hassett, 2018; LeClere, 2018; Manžuch, 2017).

1.3 Digital Cultural Heritage Collections

Cultural heritage refers to past legacy about how a person lives at the present time and how the aspects are passed to the future generation. Cultural heritage is divided into two categories (Abd Manaf, 2007):

i. Tangible cultural heritage refers to objects that are (a) movable (e.g. paintings, antiquities and artefacts) as well as (b) immovable (e.g. buildings, monuments and archaeological sites).

ii. Intangible cultural heritage refers to objects that cannot be touched, but which can be felt through other sensory organs, such as those that can be seen during a play or dance performance, or heard when stories are read or music is played.

Digitisation of tangible and intangible heritage objects has created a new form of cultural heritage known as digital cultural heritage or cultural heritage information resources (Lor & Britz, 2012). Cultural heritage information resources include a vast range of objects, contents and artefacts. The European Commission have asserted that the cultural memory of Europe is composed of prints (newspapers, journals, books etc.), photographs, museum objects, archival documents, sound and audio-visual materials, monuments and archaeological sites.

The three fundamental activities integral for generating and applying the digital cultural heritage are digitisation, access and preservation (Alvey, 2016). The first activity is digitisation which refers to conversion of objects into digital format from the analogue form. Nonetheless, digital objects without the analogue form need not undergo this process, but it is replaced by object creation step. The second activity is providing access to digital heritage in which users can view the object apart from having intuitive and efficient tools to seek resources. The last activity which refers to preservation, ascertains the availability and continuous function of a digital object at present and in future. Cultural heritage objects which differ from published objects in libraries that can be found in other places, are unique with limited accessibility and usage in their original form physically. Nonetheless, since digitised materials have no geographic boundaries, they bridge the cultural variance gap and are viable for educational purposes.

Upon digitisation of cultural heritage objects, metadata have a significant role as they enhances the usability and efficacy of search systems by offering a range of access points, preserving both aspects of contexts and semantics as well as linking similar materials that have multiple versions with those from similar collections (Schlötterer *et al.*, 2014). Metadata offer detailed and general frameworks for specific community and for resource search across varied communities. Additionally, metadata include right's management and preservation of information.

It has been suggested that adjustments need to be done by cultural heritage institutions on their digital collection planning to suit the nature of the enlisted objects and the needs of users. Cultural collections normally are rich with unique features and inclusive of physical objects, written texts, maps, photographs, sound recordings, and in some cases, original digitised objects (Schlötterer *et al.*, 2014). Therefore, it is a norm to stumble upon collections that are rich in semantics and intricate structures. It is better to separate the composite objects (e.g. traditional costumes or photographs with texts) in cultural heritage collections into parts based on their structures to characterise them individually with suitable metadata elements. In fact, digitisation has been regarded as the best solution by cultural heritage institutions to preserve vulnerable and rare objects. Thus, digitisation practices must be adhered to when compiling digital collections, largely to preserve all data concerning the digitisation devices and processes.

Flexible and rich metadata models are needed to manipulate and represent cultural information objects that are not only complex, but also with intense heterogeneity. Furthermore, high quality metadata content is required to retrieve content in a more effective manner. Poor metadata content generates irrelevant outcomes. Hence, an effective retrieval framework is sought to address issues related to low-quality metadata contents especially in digital cultural heritage collections.

1.4 Information Retrieval

Conventionally, the IR systems retrieve information from unstructured texts. IR chooses and ranks documents based on the likelihood of relevance to the needs of the users. IR traces relevant data in line with the query submitted by a user. The IR systems have turned omnipresent primarily because content search via high-speed networks appears to be the most viable solution in scavenging relevant data from the rapidly growing digitised unstructured information in massive amount. IR incorporates two expansion methods (document expansion (DE) and query expansion (QE) methods) as well as basic IR models (vector space, Boolean, probabilistic and language models).

1.4.1 Information Retrieval Expansion Methods

Different expansion methods are applied to enhance document retrieval such as by embedding new words that may appear in documents relevant to the user‘s query.

The two primary methods of expansion in IR are: (i) document expansion (DE) (expansion at index time) and (ii) query expansion (QE) (expansion at query time).

1.4.1(a) Document Expansion Method

Instead of expanding a certain query from the vast retrieved documents that takes up much time, the DE method expands documents with potential terms in the query that is found in the same type of documents. DE is performed before indexing time and it is costly. The time for query is increased slightly as the average inverted lists maybe 10% longer based on the selected DE approach (Billerbeck & Zobel, 2005). All the methods of expanding documents share one goal - to discard unproductive run-time QE. Since the IR approach involves expansion methods, it can be improved by enhancing the retrieval performance. Besides, these methods are potential candidates for enhancing both quality and retrieval issues of the metadata content.

1.4.1(b) Query Expansion Method

Numerous methods can be applied to expand a user's query for retrieval of relevant documents. Despite on the fact that some search engines may function well to some degree, glitches still occur, for instance mismatch of words issue, as most of the IR systems compare the terms between document and query on lexical level instead of semantic. In fact, Wen *et al.* (2002) asserted that the length of queries on average is about two or three words. Incompatible document and query terms as well as short queries, can heavily affect the process of retrieving relevant documents. The QE methods can enhance IR effectivity by incorporating extra phrases or terms to the query so as to enhance document retrieval. Selecting the expansion terms from the original query is a major issue in QE. The two QE techniques are (Soni & Singh, 2014): (i) global analysis technique and (ii) local analysis technique.

In the global analysis technique (Han & Chen, 2009), the query is expanded based on the extracted information from all the documents found in the database. Clustering which is the earliest global technique (Soni & Singh, 2014), groups the terms found in a document into clusters based on frequency to expand the user's query. Several other global analysis methods are Latent Semantic Indexing, Similarity Thesauri, and Phrase Finder (Park & Ramamohanarao, 2007). These global analysis techniques demand statistical analyses (e.g. statistics on paired terms co-occurrences) that generate similarity matrix between the terms. In expanding the query, terms that closely resemble the query are detected and incorporated to the original query. As for the local analysis technique, expansion of query is performed by embedding expansion words from documents that are found to be relevant (Alhroob *et al.*, 2013). Initially, a query is submitted and the IR process is performed. Next, documents that are ranked high are assessed for QE. As for the local strategy, high-ranked documents are assessed in order to identify suitable terms meant for query expansion.

1.4.2 Information Retrieval Models

Various models are found in the IR literature to suit a range of notions in line with retrieving documents relevant to query. Whilst certain models (e.g. Boolean model) have been significant in the past, the present form of IR is ranked by the ability to retrieve. According to Christopher *et al.* (2008), query reflects a group of keywords that are not in order which is also termed as 'bag of words'.

In statistically determining the distribution of terms in an array of collections and documents, a similarity measure between document and query is calculated by the IR system in which a document list is provided in order of decreasing similarity score (Büttcher *et al.*, 2016). In fact, varied models determine document-query similarities in differing ways. The three most popular and successful families of models (Sanz-Cruzado *et al.*, 2018) are vector space, probabilistic and language models. This work focuses on the language model due to its ability in outperforming the other models for enhancing retrieval issues of the metadata content, as discussed in Chapter Two.

1.4.2(a) Language Model

Language model estimates word dispersion in the input language. Based on the IR context, a document reflects a sample that is derived from a language model. Document refers to one channel of conveying information with terms in the vast collection generated by certain likelihood. Rank of documents is determined by the probability that the terms in the user's query are generated by document language model. Some language models in light of IR include multiple Bernoulli models (Feng & Manmatha, 2008), multinomial models (Zhou & Liu, 2008) and relevance models(Ogilvie & Callan, 2003a). Regardless of the variations in model applications, the related processes are composed of these three primary steps: (1) each document in the collection has an estimated language model, (2) the system calculates the probability that we would observe the sequence of query terms if we sample terms at random from each document language model, and lastly, (3) the documents are ranked in order of these probabilities.

1.5 Motivation

The rapid growth of DROs (e.g. in cultural heritage domain) and the valuable content in such resources have increased the accessibility of these resources to the users. Nonetheless, in attempting to enhance accessibility to resources, rich metadata content is required to cater the needs of users as well as to provide search outcomes closer to those requested. Digital materials pose many challenges in retrieving their content to interested users. The challenges include low metadata quality that makes the search process less effective especially when the user is not an expert in the domain area. The main motivation of this research derives from the need for a more effective IR system that enriches and handles metadata content for non-expert users. Hence, there is a need for better and effective methods and models that can be incorporated into IR so as to allow a user to easily access and explore the information on DROs especially cultural heritage collections.

1.6 Problem Statement

Metadata unit is the only way to express DROs especially in cultural heritage collections. Metadata appear to be the main factor that directly influences the DRO search (Cechinel *et al.*, 2009). Due to the fact that metadata quality mainly affects the effectiveness of the search process, many researchers such as Ternier *et al.* (2009), and Bui and Park (2013), have assessed the impact of metadata quality on the search result of DROs and revealed that the deficiencies of metadata quality include lack of metadata content of DROs and generation of irrelevant outcomes in the search process. Additionally, Gaona-García *et al.* (2017) and Seifert *et al.* (2017) verified the lack of metadata content quality of DROs and asserted that there are still rooms for further expansions and enrichment in order to improve access and quality of metadata content. In fact, a substantial number of researchers have addressed the deficiencies found in the quality of the metadata content. The first is to solve the issue by treating DROs as databases and to retrieve them through data retrieval (Kettula & Hyvönen, 2012; Zervanou *et al.*, 2011).

The recent research trend treats IR components individually to handle DROs such as document expansion method (Kando & Adachi, 2004; Min *et al.*, 2010), retrieval model (Tan, 2015; Tan & Lim, 2017) and query expansion method (Akasereh, 2013b; Almasri *et al.*, 2014b; Kanhabua *et al.*, 2016). These researchers have proven that individual IR components can improve the performance of the DRO retrieval. However, they have excluded the nature of metadata and related issues. They seemed to have treated DROs directly as unstructured text which lead to reduction of the effectiveness and zero probability problem in the DRO retrieval results. Nevertheless, Koolen *et al.* (2007b) revealed that conventional IR techniques cannot be straightforwardly applied on DROs. Therefore, there is a need to enhance the DRO retrieval performance by improving individual IR components as well as taking into account the nature of metadata and related issues.

1.7 Research Objectives

The main goal of this research is to overcome the problems of the existing DRO retrieval (such as lack of quality of metadata content and difficulty in accessing metadata content) by proposing an IR framework. This can be further subdivided into the following objectives:

i. To enhance the document expansion (DE) method by feeding and providing metadata content with new information to increase the effectiveness of the DRO retrieval.

ii. To formulate the Dirichlet smoothing (DS) model that is able to retrieve the most relevant results (documents or metadata units) related to a particular query in order to improve the matching process in DRO retrieval.

iii. To enhance the query expansion (QE) method by improving the quality of candidate terms to be added semantically to the user's query in order to improve the performance of DRO retrieval.

iv. To design and integrate the above-enhanced expansion methods and the proposed retrieval model for a more effective DRO retrieval in light of cultural heritage collections.

1.8 Research Contributions

The contributions of this research are as follows:

i. An enhanced document expansion (EDE) method that utilizes DRO documents and feed documents with new information in order to increase the effectiveness of the DRO documents. The EDE method improves the weight measurement for metadata unit terms, improves the probability estimation equation, and introduces a new search space called backup.

ii. An adaptive structured Dirichlet smoothing (ASDS) model that improves the matching assumption in the Dirichlet smoothing (DS) model. The ASDS model involves two sub-models. The first model is structured Dirichlet smoothing (SDS) which is concerned with the document structure while the second model is the adaptive Dirichlet smoothing (ADS) model that adapts the parameters used in the DS model.

iii. A semantic query expansion (SQE) method that improves the performance of DRO retrieval by improving the quality of candidate terms to be added semantically to the entire query terms.

iv. An IR framework (IRF) (two-stage expansion method and a retrieval model) for DROs retrieval that addresses problems related to non-expert users such as short query and short document problems.

1.9 Scope and Limitations

The application domain of this research is DRO. Europeana cultural heritage collection CHiC2013 (Documents in English) which provides 22 test queries will be used as a dataset for this research. Therefore, the evaluations carried out in this research have to be restricted to 22 test queries and documents in English. There are a number of IR performance measurements. However in this research, Mean average precision (MAP), precision at ten (p@10) and Precession-Recall measurements are used for the evaluations because the benchmarks for this research presented their results using these measurements. Also these benchmarks used the same dataset (CHiC 2013) for evaluations.

This research focuses on improving the retrieval performance of DRO documents and metadata by using only IR components (methods and retrieval models) in order to solve the lack of quality in DRO metadata contents and the difficulty in accessing DROs. So, this research will not deal with another retrieval approach which is the DR approach. This improvement will help a non-expert user to access the DRO contents. Therefore, this research will look into enriching metadata content, improving access to the user by expanding user's query, and enhancing the retrieval models to reduce the zero probabilities of DRO retrieval. Although heterogeneity problem is prevalent in DROs and metadata and it may affect DRO retrieval performance, the problem can only resolved by the DR approach. Therefore, the DRO heterogeneity problem is beyond the scope of this thesis.

1.10 Thesis Organisation

Chapter 2 presents a brief overview on data retrieval for DROs along with their issues and limitations. This chapter provides a detailed overview of IR particularly DE and QE methods as well as language models. It discusses a detailed overview of the language model inclusive of its gram models, smoothing models and estimation models as well as a comparison that involves advantages, disadvantages and usage of the models. Finally, a number of studies pertaining to data retrieval and IR are systematically analysed in terms of performance and the possibility of working with DRO particularly for cultural heritage collections. The research methodology, the description of CHiC2013 collection and the evaluation procedures applied in this thesis are elaborated in Chapter 3.

In Chapter 4, a new EDE method is presented to enrich the DRO’s content by solving DROs short content problem that exerted a negative impact on DRO retrieval performance. Three versions of DS models are proposed in Chapter 5 to address issues concerning DS model with DROs such as fixed smoothing parameter value problem, entire document retrieval problem, and zero probability problem.

The SQE method is proposed in Chapter 6 to solve the short user query problem by enhancing the quality of candidate terms to be incorporated semantically to the entire query terms. The integration of the proposed correlation algorithm relies on simple sensible Boolean heuristics and Wikipedia as the external resource.

In Chapter 7, several IRF versions are introduced to enhance the performance of DROs retrieval by improving the quality of retrieved documents contents, improving the retrieval model and improving the user’s query. The principle task of IRFs is to make all components of the IR expansion methods (DE and QE) and retrieval model (DS) work together to achieve the greatest benefit in improving the retrieval performance. Lastly, Chapter 8 concludes the thesis and highlights several ideas for future research.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

The chapter elaborates the literature review of various related work pertaining to the performance of the digital recourse object (DRO) retrieval. Recently, the need for enriching and accessing DROs has been addressed differently by both the data retrieval (DR) and information retrieval (IR) research communities. A range of methods meant to enrich and retrieve DROs using DR has been proposed. Apart from that, many available approaches that can be implemented to retrieve and enrich DROs using IR and their limitations are discussed. Despite the vast studies concerning DRO's access, only a handful have addressed problems related to DROs (Aruleba *et al*., 2016; Ghorab *et al.*, 2013; Haslhofer & Klas, 2010). Hence, this literature review is the first attempt to cover and compare enrichment of DRO's accessibility based on dual ways, namely DR with three branches (mapping generic schema, mappings across existing schemas, and mappings of existing information to ontology) and IR with two branches (expansion methods and retrieval models).

This chapter is structured as follows: Section 2.2 discusses DR and its related works while Section 2.3 presents IR (expansion methods and retrieval models). Next, Section 2.4 depicts the comparison between IR and DR whereas Section 2.5 discloses prior analyses and findings. Lastly, this chapter is concluded in Section 2.6. The overview of the literature review for this study is illustrated in Figure 2.1.

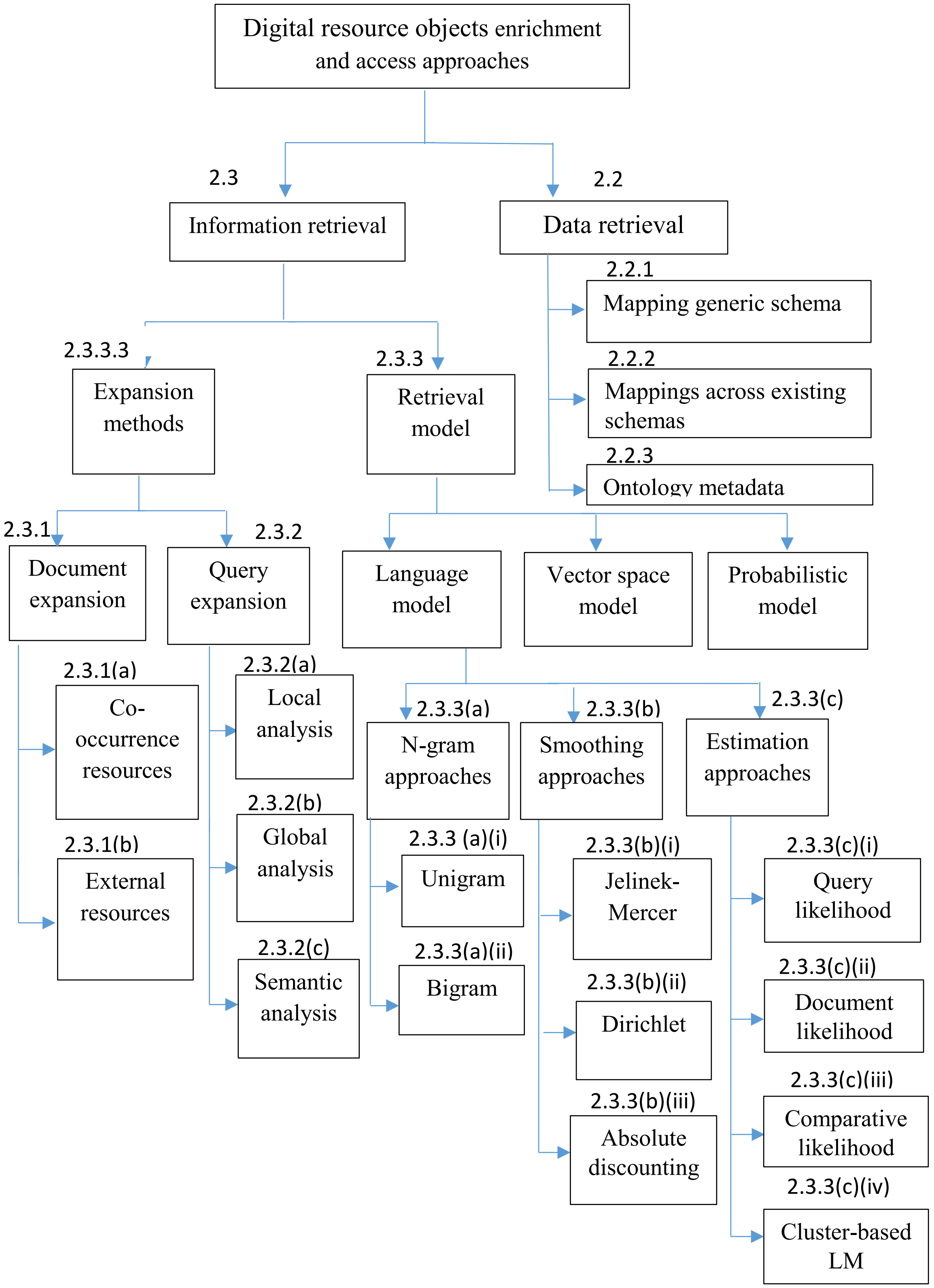

Figure 2.1: An overview of the literature review

2.2 Data Retrieval

In DR, exact matches are sought using stringent queries on structured data, such as DROs. DROs use metadata as data schemas or elements to describe the contents. Various metadata standards can describe a range of objects. Several researches (Kando & Adachi, 2004; Kanhabua et al., 2016; Signore, 2008) asserted that retrieval of DROs under DR mainly involves the following issues:

i. Lack of quality in DRO metadata contents (Hampson *et al.*, 2012; Manguinhas *et al.*, 2016; Zervanou et al., 2011): The content of metadata in DRO is insufficient for the user and needs to be enriched with more information (Agosti *et al.*, 2013; Álvarez, 2015; Berardi *et al.*, 2012; de Boer *et al.*, 2013). A user may face problem in retrieving adequate and fruitful information from free text documents, and this problem is called short document problem. Sokvitne (2000), who assessed the effectiveness of Dublin Core Metadata Initiative (DCMI) for retrieval, reported that metadata contents are nearly useless for retrieval due to their poor quality. In order to enrich DRO's content for cultural heritage (CH) collection, Agirre *et al.* (2012) developed a small dataset comprising of 400 objects from European Digital Library (Europeana) that were manually linked to Wikipedia whenever there exists an article that exactly describes the same object as the CH object. The dataset was used to evaluate two systems that yielded relatively poor performances. An enrichment framework proposed by Gavrilis *et al.* (2015) for content reuse was aggregated, transformed, and enriched for Europeana in tourism domain to automate this task. Seifert et al. (2017) claimed that lack of metadata content quality in DROs leads to failure in retrieving relevant documents.

ii. Difficulty in accessing DROs: The limited knowledge of a user on the metadata content of DROs makes accessing and exploring the available metadata content difficult (Elsweiler *et al.*, 2010; Walsh *et al.*, 2018; Windhager *et al.*, 2018). In DROs, such as in CH collection, the users, especially non-experts, deal with a domain-specific collection with its specific terminology wherein the collection is searched by many users who do not necessarily use specific terminology in their queries, and hence making the matching process more difficult (Parikh *et al.*, 2013). In order to make DROs more accessible and user-friendly, a system that incorporates three components, namely personal metadata, semantic enrichment and query answering techniques was formulated by Kollia *et al.* (2012). In another work, open data sources as proposed by Tonkin and Tourte (2016), offered a landscape to understand and interpret metadata content in DROs.

iii. The heterogeneity of metadata content in DROs: Data are distributed over different schemas of databases and users cannot seek information due to unfamiliarity with these schemas. Heterogeneity at the level of schemas makes accessibility difficult (Candela *et al.*, 2007). Early studies of DROs focused on architecture of single collection to solve heterogeneity issues (Arms, 1995; Barros *et al.*, 2008; Best *et al.*, 2007; Peterson *et al.*, 2003). A semantic relation of interoperability of heterogeneous DROs was weighed in in several studies (Liao *et al.*, 2010; Qi *et al.*, 2005). Many frameworks (Ikonomov *et al.*, 2013; Windhager *et al.*, 2016) and systems (Brocks *et al.*, 2001; Hampson et al., 2012; Seifert et al., 2017) have been developed to offer access to heterogeneous metadata schemas in DROs. In DROs, making its contents more accessible with rich metadata content and interoperability is an active research area, especially for CH collections that consist of objects described by a variety of metadata schemas. Mapping schemas are often used to achieve metadata integration which is a crucial aspect (Madhavan *et al.*, 2001). The mapping schemas can be classified into three categories: (i) Mapping generic schema, (ii) Mapping across existing schemas, and (iii) Ontology mapping. Many efforts have been put towards the development and enrichment of metadata schemas in DROs, so as to handle them as a database.

2.2.1 Mapping Generic Schema

A schema that contains all information about any domain on the metadata for integration is called generic schema and it serves as a common interface for querying all heterogeneous metadata. Upon identifying a generic schema, manual mapping is performed to extract the existing mappings between the metadata. In addressing metadata issues across varied types of object collection, generic metadata schemas have been proposed by many researchers (Kollia *et al.*, 2012; Madhavan *et al.*, 2001; Orgel *et al.*, 2015). The elements incorporated into the proposed schema in CH by Lilis *et al.* (2005) include refined term table of contents, schema with elements related to all the corresponding terms, and element date with all the terms, mainly to enrich the collection description with time-dependent information. Several heterogeneous schemas and a generic schema were formulated by Rebaï *et al.* (2015) whereby the matching process was performed between two schemas to determine the existing mappings.

The manual mapping that extracts the existing mappings between the metadata was a failure due to inefficient cost and time. The main limitations of generic schemas as reported by Gergatsoulis *et al.* (2010) and Bellini *et al.* (2010), are deficiency and lack in definition of CH collections. Peroni *et al.* (2013) asserted that generic schemas, such as DCMI, have element levels with poor organisation of the hierarchal structure and have extensive use of element levels with generic terms.

2.2.2 Mappings across Existing Schemas

Many researchers have integrated diverse metadata schemas via mappings across existing schemas. Thus, an existing schema is applied among multiple metadata schemas although there is no direct mapping between metadata schemas. This allows incorporation of mappings from diverse metadata schemas (Godby *et al.*, 2003). The DCMI (Baker, 2012; Initiative, 2012; Johnston & Powell, 2010) and the Simple Knowledge Organisation System (SKOS) (Miles & Bechhofer, 2009) are commonly employed as the existing schemas.

Several experimental case studies on mappings across existing schemas carried out by Van Gendt *et al.* (2006) and Hyvönen *et al.* (2016) revealed that the available schema models are better than the newly-created schemas. This model was employed to enrich DROs such as in the work by Zervanou *et al.* (2011) who addressed the poor quality of metadata contents in CH domain. The study proposed enrichment of the existing CH metadata with automatically generated semantic metadata content. The terms recognised in the metadata content are used to indicate the metadata content. The inter-relationships among the recognised terms in the hierarchy are assumed to reveal the knowledge structure of the CH documents.

Similarly, Gavrilis *et al.* (2015) introduced the idea of re-using aggregated and enriched CH metadata content that composed of a generic enrichment service that offers a series of enrichment. Enrichment enriches each metadata schema and hence, enrichment supports specific metadata schemas. Berardi *et al.* (2012) discovered that increasing metadata quality content leads to more effective content retrieval. El-Sappagh *et al.* (2011) claimed that transferring metadata content across multiple schemas leads to information loss during transformation from one metadata schema to another.

2.2.3 Ontology Mapping Schemas

Ontology mapping has a role in semantic interoperability scenarios. It expresses semantics in a formal manner which can be used as an umbrella of terms and meanings to express the same subject or concept (Partridge, 2002; Uschold & Gruninger, 1996). The most common ontologies in DROs are Europeana Data Model (EDM) and Conceptual Reference Model (CRM**)** mediator ontology (Doerr, 2003). Ontology refers to a conceptual representation that ascertains semantic integration between varied DRO metadata schemas apart from discarding potential semantic heterogeneities. Studies that have addressed heterogeneity in CH collections using ontologies mediator were undertaken by Kakali *et al.* (2007*),* Lourdi *et al.* (2009) and Tomasi *et al.* (2015).

Two ontologies were developed by Daquino *et al.* (2017), one to represent the political roles of agents in an event-centric perspective, and the other to manage interpretation and served as individual hermeneutical approaches to the document content. Ontologies offer solutions to semantic heterogeneity issue. A semantic ontology developed by Carrasco (2013) promotes semantic integration between different metadata schemas to integrate CH metadata contents. Hajmoosaei and Skoric (2016) devised a top-level ontology based on metadata and local ontology for archive data source. They reported that the ontology is an effective way to solve heterogeneity issue, but it demands expert knowledge and proper development methodology. The ontology makes CH collections difficult to manage, thus leading to intricacy in schema mapping (Wache *et al.* (2001). Despite its automatic transformation, Tallerås *et al.* (2014) reported that ontology data require manual assistance for quality check. Table 2.1 presents the limitations of DR approaches while Table 2.2 summarises prior studies on DR in DROs, as reviewed in this section.

Table 2.1: Limitations of DR mapping approaches in DROs

Approach	**Limitation**	**Reference**
Mapping generic Schema	• Deficient and lacks definition in cultural heritage collections	Gergatsoulis *et al.* (2010)
	• Element levels are organised without strong hierarchical structure • Uses extensive element levels with generic terms	Peroni *et al.* (2013)
	• Costly and time consuming • Uses manual mapping	Rebai *et al.* (2015)
Mappings across existing schemas	• Ignores semantic heterogeneity problem	Gendt *et al.* (2006)
	• Difficult to maintain especially with updates by different users	Hyvonen *et al.* (2016)
	• Loses information during transformation from one metadata schema to another	El-Sappagh *et al.* (2011)
Ontology mapping schemas	• Data are transformed automatically, but demands human handling	Talleras *et al.* (2014)
	• Ontology makes cultural heritage difficult to manage, leading to intricacy in schema mapping	Wache *et al.* (2001); Kakali *et al.* (2007)
	• Ontology needs to be created with expert knowledge and proper development methodology	Hajmoosaei and Skoric (2016)

Table 2.2: Summary of DR in DROs

Authors (Year)	Research Domain	Approaches	Issues	Findings
Lilis *et al.* (2005)	Cultural heritage	Mapping generic Schema	Lack of metadata content	Multidimensional metadata model that enriches a metadata application profile
Gendt *et al.* (2006)	Cultural heritage	Mappings across existing schemas	Heterogeneous metadata	A prototype for semantic web techniques to match the vocabulary of the collection.
Lourdi *et al.* (2009)	Cultural heritage	Ontology mapping schemas	Heterogeneous metadata	Crosswalk between Dublin Core Collections is presented
Zervanou *et al.* (2011)	Cultural heritage	Mappings across existing schemas	Lack of metadata content	A methodology for semantically enriching archival description metadata
Peroni and Shotton (2012)	Publishing domain	Mapping generic Schema	Lack of metadata content	Ontology for describing bibliographic resources and bibliographic citations on the Semantic Web
Kollia *et al.* (2012)	Cultural heritage	Mapping generic Schema	Metadata access	A semantic query answering approach that assists content providers and users to enrich their data
Carrasco *et al.* (2013)	Cultural heritage	Ontology mapping schemas	Heterogeneous metadata	A semantic mapping that provides interoperability amongst the cultural heritage systems
Tomasi *et al.* (2015)	Cultural heritage	Ontology mapping schemas	Heterogeneous metadata	A model of multiple ontologies to explore the semantic content of heterogeneous digital collections
Gavrilis *et al.* (2015)	Cultural heritage	Mappings across existing schemas	Lack of metadata content	A framework for re-using aggregated and enriched metadata for cultural heritage related content for the tourism industry
Rebaï *et al.* (2015)	XML French collection	Mapping generic Schema	Heterogeneous metadata	A matching process that deals with the metadata schema heterogeneous