

**DIGITAL QURAN WITH STORAGE OPTIMIZATION  
THROUGH DUPLICATION HANDLING AND COMPRESSED  
SPARSE MATRIX METHOD**

Ashraf Saleh Mohammad Alomoush

Thesis submitted in partial fulfilment for the degree of  
DOCTOR OF PHILOSOPHY  
SCIENCE AND TECHNOLOGY

UNIVERSITI SAINS ISLAM MALAYSIA

November 2022

## AUTHOR DECLARATION

I hereby declare that the work in this thesis is my own except for quotation and summaries, which have been duly acknowledged.

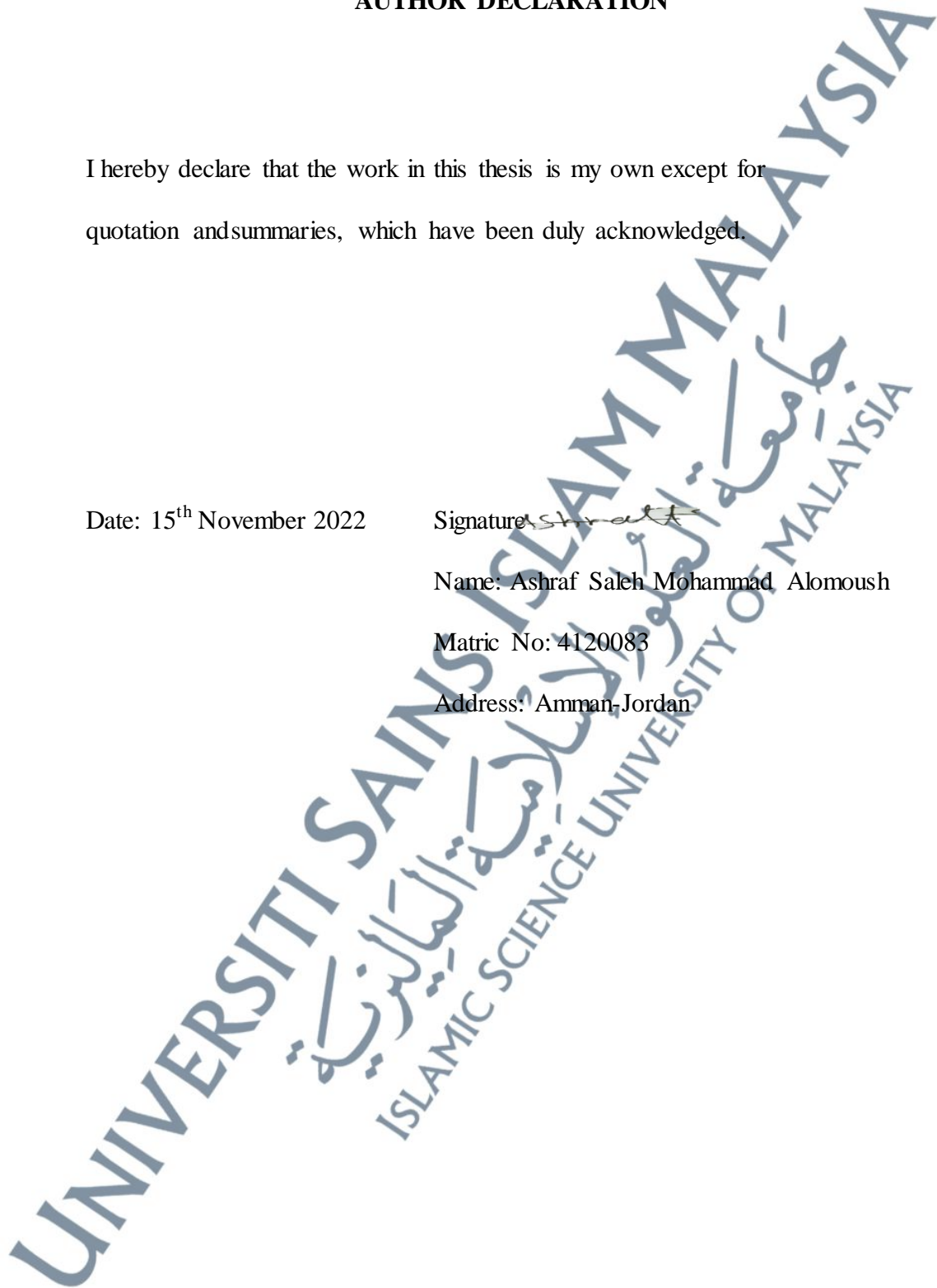
Date: 15<sup>th</sup> November 2022

Signature: 

Name: Ashraf Saleh Mohammad Alomoush

Matric No: 4120083

Address: Amman, Jordan



## ACKNOWLEDGEMENTS

Praise is to ALLAH, the Almighty, Most Gracious, and Most Merciful, who enabled me and gave me everything to complete this thesis. I would also like to extend my thanks and appreciation to my supervisor, Prof. Ts. Dr Norita Md Norwawi, who was the source of inspiration at all stages of this research. Much of this work would have been impossible without her guidance, comments, and encouragement.

In the life of Saleh, the father

In the life of Aziza, the mother

To my wife

To Darling Dawood, the son

To Darling Hor, the daughter

To Darling Kenz, the daughter

To my friends

May ALLAH bless you all.....

## ABSTRAK

Penyelidikan ini mengkaji topik Al-Quran Digital sebagai aplikasi yang digunakan oleh sebilangan besar orang di seluruh dunia dan melalui pelbagai peranti (cth. telefon pintar, tablet dan komputer). Ketika ini Quran Digital berformatkan teks, memanfaatkan perwakilan Hexadesimal. Namun, ruang storan tidak dioptimumkan oleh penggunaan penyatuan rentetan aksara dalam Unicode bagi suatu kalimah Quran. Maka, saiz storan akan berkadar terus dengan panjang kalimah apabila ditukar bentuk Unicode. Oleh kerana itu kajian ini mencadangkan model Al-Quran Digital (DQM) yang baharu bertujuan untuk mengurangkan saiz ruang penyimpanan melalui perwakilan kalimah dalam Hexadecimal menggantikan penyatuan rentetan aksara. Selanjutnya saiz storan dioptimumkan lagi dengan pengendalian kalimah Quran yang berulang. Ini selari dengan struktur matriks jarang yang dimampatkan untuk perwakilan ayat-ayat dan surah-surah Al-Quran. Fokus penyelidikan ini adalah pengoptimuman storan untuk Quran Digital menggunakan format teks. DQM telah dibangunkan menggunakan Visual Studio dan pelayan Java dimana kualiti penyelesaian diukur dengan perbandingan saiz fail sebelum dan selepas menggunakan model DQM. Untuk surah Al-Baqarah, pengurangan saiz storan ialah 50.76%, manakala surah Al-Fatihah ialah 69.81%. Hasil penyelidikan ini konsisten dengan kajian lain dalam konteks yang sama dan memberikan implikasi dalam kedua-dua domain teori dan praktikal. Selain dari analisis pengurangan saiz ruang, pakar bidang al-Quran juga ditemuduga berkaitan Al-Quran Digital yang dibangunkan. Hasil kajian ini membantu pembangun aplikasi membina Quran Digital yang boleh dimanfaatkan oleh pengguna kerana kebolehpercayaan, kesahihan dan pengoptimuman storan supaya ia boleh digunakan sebagai aplikasi standard.

## ABSTRACT

This research looks into the topic of the Digital Quran as an application used by many people worldwide and via various devices (e.g. smartphones, tablets, and computers). The current digital Quran text-based format benefitted the Hexadecimal representation. However, space was not optimized due to the string concatenation approach of the Unicode of each letter that occurs in a word. Thus, the storage size for the Unicode representation is directly proportional with the length of a word. In this regard, the current study proposes a new Digital Quran Model (DQM) that aims to reduce storage requirements through word conversion into Hexadecimal instead of combining letter sequences. This approach is further optimized through Quranic word duplication handling. This is paralleled with a compressed sparse matrix representing Quranic verses and surah. The focus of this research is storage optimization for text-based Quran content formats. DQM was implemented using Visual Studio and Java servers, where the solution quality was measured by comparing the file size before and after applying the DQM model. For surah Al-Baqarah, the reduction in the storage size was 50.76%, and Al-Fatihah was 69.81%. The results of this research are consistent with other studies in the same context and provide implications in both theoretical and practical domains. In addition to space reduction analysis, the current research interviews domain experts on the proposed digital Quran applications. The outcome helps the researchers to develop a Digital Quran that users can benefit from due to its reliability, validity, and optimized storage so that it can be used as a standard application.

## المخلص

يبحث هذا البحث في موضوع القرآن الرقمي كتطبيق يستخدمه العديد من الأشخاص في جميع أنحاء العالم وعبر الأجهزة المختلفة (مثل الهواتف الذكية والأجهزة اللوحية وأجهزة الكمبيوتر). أستفاد التنسيق الرقمي الحالي المستند إلى النص القرآني من التمثيل السداسي العشري. ومع ذلك، لم يتم تحسين المساحة بسبب نهج تجميع السلسلة لتشفير كل حرف يحدث في الكلمة الواحدة. وبالتالي، فإن حجم التخزين لتمثيل اليونيكود يتناسب طردياً مع طول الكلمة. في هذا الصدد، تقترح الدراسة الحالية نموذجاً جديداً للقرآن الرقمي يهدف إلى تقليل متطلبات التخزين من خلال تحويل الكلمات إلى نظام سداسي عشري بدلاً من الجمع بين تسلسل الحروف. تم تحسين هذا النهج بشكل أكبر من خلال تقليل التكرار للكلمات و يتوازي هذا مع مصفوفة متفرقة مضغوطة تمثل الآيات والسور القرآنية. يركز هذا البحث على تحسين التخزين لتنسيقات محتوى القرآن النصية. تم تنفيذ نموذج القرآن الرقمي باستخدام البرمجة المرئية وخوادم الجافا حيث تم قياس جودة الحل من خلال مقارنة حجم الملف قبل وبعد تطبيق نموذج القرآن الرقمي. بالنسبة لسورة البقرة، كان الانخفاض في حجم التخزين ٥٠.٧٦٪، والفاصلة ٦٩.٨١٪. تتوافق نتائج هذا البحث مع دراسات أخرى في نفس السياق وتوفر آثاراً في كل من المجالات النظرية والعملية. بالإضافة إلى تحليل تقليل المساحة، يقوم البحث الحالي بإجراء المقابلات مع خبراء المجال حول تطبيقات القرآن الرقمي المقترحة. تساعد النتيجة الباحثين على تطوير مصحف رقمي يمكن للمستخدمين الاستفادة منه نظراً لموثوقيته وصلاحيته وتخزينه الأمثل بحيث يمكن استخدامه كتطبيق قياسي.

## TABLE OF CONTENTS

CONTENT	PAGE
AUTHOR DECLARATION	i
ACKNOWLEDGEMENTS	ii
ABSTRAK	iii
ABSTRACT	iv
AL-MULAKHKHAS	v
LIST OF TABLES	ix
LIST OF FIGURES	xi
LIST OF APPENDICES	xiv
LIST OF ABBREVIATION	xv
CHAPTER 1: INTRODUCTION	
1.1 Introduction	1
1.2 History of Quran Printing and Publications	1
1.3 Problem Statement	4
1.4 Research Questions	7
1.5 Research Objectives	7
1.6 Research Scope and Limitation	8
1.7 Research Contributions	9
1.8 Organization of the Thesis	10
1.9 Summary	11
CHAPTER 2: LITERATURE REVIEW	
2.1 Arabic Language Foundation	13
2.2 Quran Publications	15
2.3 Digital Quran	18
2.4 Digital Quran Model Development	21
2.5 Digital Quran Publications	23
2.5.1 Quran Text-Based Format	28
2.5.2 Quran Image-Based Format	33
2.5.3 Quran Audio / Video-Based Format	36
2.6 Vulnerability Issues for Digital Quran	44
2.7 Digital Quran Content Integrity	46
2.8 Related Works	49
2.9 Character Representation for Arabic Letters	56
2.10 Representation of Words and Verses in the Quran	59
2.11 Data Compression Using Hexadecimal	60
2.12 Discussion	61
2.13 Summary	62

### CHAPTER 3: RESEARCH METHODOLOGY

3.1	Introduction	63
3.2	Proposed Digital Quran Text Format Using Hexadecimal and Compressed Matrix Structure	65
3.2.1	Arabic Letters and Words Conversion to Hexadecimal Representation	66
3.2.2	Compressed Matrix Representation for Quran Content	69
3.2.3	Sparse Matrix with Double Offset Indexing	72
3.2.4	Handling Word Duplication of the Digital Quran Content Structure in the Matrix Representations	73
3.3	Development Phase: Digital Quran Model Implementation	75
3.3.1	Creating Quran Test case Data: Surah Al-Fatihah and Al-Baqarah	75
3.3.2	Consultation with Domain Expert	76
3.3.3	Design and Develop the Digital Quran Proof of Concept	76
3.4	DQM Evaluation Phase	80
3.5	Summary	82

### CHAPTER 4: NEW DIGITAL QURAN MODEL TO OPTIMIZE STORAGE

4.1	Introduction	83
4.2	Letters and Words Conversion	84
4.2.1	Quran Words Conversion Algorithm	85
4.2.2	Quran Verses Conversion for Surah Al-Baqarah	88
4.3	Storage Optimization with Compressed Content Structure and Duplication Handling	89
4.4	Sparse Matrix Compression with Double off Set Indexing	92
4.5	New Digital Quran Model with Lightweight Storage	94
4.6	Discussion and Consultation	95
4.7	Summary	97

### CHAPTER 5: DIGITAL QURAN MODEL IMPLEMENTATION AND EVALUATION

5.1	Introduction	98
5.2	DQM Design and Development	98
5.2.1	DQM Prototype Development	100
5.2.2	Implementation Tools	101
5.2.3	Coding DQM	103
5.2.4	Word and Verses Conversion	104
5.3	Storage Optimization with Compressed Sparse Matric and Handling Duplications	110
5.4	Letters and Words Conversion for Surah Al-Baqarah	114
5.5	Surah Al-Fatihah Implementation and Performance Evaluation	115
5.6	Surah Al-Baqarah Implementation and Performance Evaluation	118
5.7	DQM Evaluation Methods	124
5.7.1	Comparison Results of DQM with Existing Applications	124

5.7.2 Content Integrity and Authentication	127
5.7.3 Experts Evaluation	130
5.8 Conclusion	131
<b>CHAPTER 6 CONCLUSION AND FUTURE WORK</b>	
6.1 Overview	133
6.2 Responses to Research Objectives	134
6.3 Contribution of this Study	138
6.4 Future Research Directions	138
<b>REFERENCES</b>	<b>140</b>
<b>APPENDICES</b>	<b>149</b>

UNIVERSITI SAINS ISLAM MALAYSIA  
جامعة العلوم الإسلامية الماليزية  
ISLAMIC SCIENCE UNIVERSITY OF MALAYSIA

## LIST OF TABLES

<b>Tables</b>	<b>Page</b>
Table 2.1: Examples of Arabic Fonts and Arabic Fonts Names	18
Table 2.2: Arabic Character Representation Unicode	13
Table 2.3: Methods Applied in Tabular Format in Conjunction with The Purpose of Each Method, the Principal Method Used, and Final Results Realized	52
Table 2.4: Hexadecimal Value for Each Character	61
Table 2.5: Compressing a Quranic Verse Using Proposed Compression Method	61
Table 3.1: Comparison between letter conversion and word conversion in terms of storage.	69
Table 3.2: Evaluation of the Storage Optimization	81
Table 4.1: Statistics of the Holy Quran	84
Table 4.2: Examples of Repetition of Quran Words	85
Table 4.3: Storage Comparison Between words in Arabic and Hexadecimal	87
Table 4.4: Lookup Table to Calculate Words with Total Letters Count, Representation in Hexadecimal and Unique ID	89
Table 5.1: Comparison of Memory Size Arabic Verse and Its Hexadecimal Representation	108
Table 5.2: Comparison of Various Digital Representations for Surah Al-Fatihah	109
Table 5.3: The Most Frequently Words in Surah Al-Baqarah with Their ID's	112
Table 5.4: Sparse Code for the Longest Ayah in the Holy Quran (Al-Baqarah, 282)	113

Table 5.5: Comparatives Results for Words before and After Applying Algorithm	116
Table 5.6: Summary of the Comparison Words and Verses Conversion	117
Table 5.7: Sparse Matrix Technique to Reduce Storage Surah Al-Fatihah	117
Table 5.8: Sparse Matrix Technique with Double-off Set Indexing Compression Algorithm	118
Table 5.9: Conversion of Arabic Word into Hexadecimal Form	120
Table 5.10: Sparse Matrix Technique to Reduce Storage	121
Table 5.11: Sparse Matrix Technique & Double-off Set Indexing Compression Algorithm	122
Table 5.12: Comparison of Storage Size According to Representation: Al-Fatihah & Al-Baqarah	124
Table 5.13: Similarities and Differences in the Current Application for Digital Quran.	126

## LIST OF FIGURES

Figure	Page
Figure 2.1: Outlines the Structure of Chapter 2	12
Figure 2.2: Surah Al-Fatihah without Dots Rasm Uthmani	16
Figure 2.3: Rasm Uthmani for Surah Al Baqarah	16
Figure 2.4: Rasm Imlai for Surah Al Baqarah	17
Figure 2.5: SemQ Tool Pipeline	29
Figure 2.6: Quranic Arabic WordNet	30
Figure 2.7: Quranic Model	31
Figure 2.8: Definition of Holy Quranic Chapters	32
Figure 2.9: Quranic Images Plain and Complex	34
Figure 2.10: Watermark Encoding Process	36
Figure 2.11: Associating Quranic Verses with Web Multimedia Resources	40
Figure 2.12: Al-Anvar: Quran Research Software	41
Figure 2.13: Classification Based on Digital Quran Format	43
Figure 2.14: Unicode Centric String-Matching Approach	47
Figure 2.15: Advantages and Drawbacks of the Protection Techniques	47
Figure 2.16: Taxonomy Based on Preserving Content Integrity	48
Figure 2.17: Unicode Standard 7.0, Copyright © 1991-2014 Unicode, Inc., Arabic Presentation Forms A	57
Figure 2.18: Thad ض Letter in Four Expected Position on The Word	58
Figure 2.19: Unicode Standard 7.0, Copyright © 2014 Arabic Presentation for Thad ضLetter	58
Figure 3.1: Different Stages of the Research Activities	64

Figure 3.2:	The Unicode Standard 7.0, Copyright © 1991-2014 Unicode, Inc., Arabic Presentation Forms-A	67
Figure 3.3:	Sparse Matrix Represent Words and Verses of the Quran	70
Figure 3.4:	Sparse Matrix Populated Primarily with Zeros	71
Figure 3.5:	Sparse Matrix with Double Offset Indexing Representation	72
Figure 3.6:	Example of a Lookup Table for Indexing Unique Quran Words with Unique ID	73
Figure 3.7:	Al-Fatihah content structure with a lookup table and unique ID	74
Figure 3.8:	Phases for Digital Quran Prototype Development	77
Figure 3.9:	Comparison of Memory Space Allocated with Original Text and Hexadecimal Transformation	78
Figure 3.10:	Flowchart of the Prototype of DQM functions	79
Figure 4.1:	Outlines the Structure of Chapter 4	83
Figure 4.2:	Arabic Letters (أ, ب, ج, د, هـ, ز, ح, ط, ي) In Different Position With Their Hexadecimal Representation	86
Figure 4.3:	Word Conversion Algorithm into Hexadecimal Representation	87
Figure 4.4:	Verses Conversion Algorithm into Hexadecimal Representation	88
Figure 4.5:	DQM in Sparse Matrix with Unique ID Duplication Handling	90
Figure 4.6:	Pseudo-Code for Digital Quran Verses in Sparse Matrix with Lookup Table	91
Figure 4.7:	Double off-Set Indexing Algorithm to Reduce Storage Requirements	93
Figure 4.8:	The Algorithm Description of the Complete DQM Approach	94
Figure 4.9:	The key components of the new DQM with lightweight storage	95
Figure 5.1:	DQM General Process Flow	99
Figure 5.2:	DQM Implementation from Arabic Text Format to Hexadecimal Representation	101
Figure 5.3:	DQM Implementation Key Mechanism	102

Figure 5.4:	Anatomy Design for (قدیر) QADEER Word	105
Figure 5.5:	Implementation for (قدیر) QADEER Word	106
Figure 5.6:	Program Code Represent Surah Al-Fatihah (الفاتحة) in Hexadecimal	106
Figure 5.7:	Flow Representing the Code, Input File, Hash Map, Display and Data Set	107
Figure 5.8:	Quran Representation with Unique ID In A Sparse Matrix for Surah Al-Fatihah	110
Figure 5.9:	Comparison of Al-Fatihah Representation	111
Figure 5.10:	Screen Shot of Words Conversion (وبالوالدين, فليستجيبوا)	114
Figure 5.11:	Quran Parsing Program to Calculate Words in Hexadecimal Space Size	115
Figure 5.12:	Surah Al-Baqarah Representation	119
Figure 5.13:	Size of Arabic Text File & Hex Text for Surah Al-Baqarah	120
Figure 5.14:	Size of Arabic, Hex, and Sparse Text for Surah Al-Baqarah	121
Figure 5.15:	Arabic Text, Hex, Sparse and Double Off-Set Indexing for Surah Al-Baqarah	123
Figure 5.16:	Testing with Missed Letter Text shows that no sparse matrix was Constructed. Unmatched input text with the Look Up Table	128
Figure 5.17:	Testing with Non-Quran Text shows no sparse matrix was constructed since it does not match the Quran words in the Look-Up Table	128
Figure 5.18:	Testing with Non-Quran Text that include Words that Exist in Quran shows no sparse matrix was constructed since it does not match any Quran verses cross checked with the Look-Up Table	129

## LIST OF APPENDICES

Appendices	Page
Appendix 1: Unicode Standards	149
Appendix 2: List of Publications	151
Appendix 3: List of Awards	153
Appendix 4: Certificate of Validation	154
Appendix 5: Interview Questions	156

UNIVERSITI SAINS ISLAM MALAYSIA  
جامعة العلوم الإسلامية  
ISLAMIC SCIENCE UNIVERSITY OF MALAYSIA

## LIST OF ABBREVIATIONS

AMR	Adaptive Multi-Rate
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ASR	Automatic Speech Recognizer
DCT	Discrete Cosine Transform
HEX	Hexadecimal
DQM	Digital Quran Model
DCT	Discrete-Cosine Transform
DWT	Discrete Wavelet Transform
FFT	Fast-Fourier Transform
LSB	Least Significant Bit
PoS	Part of Speech
QoC	Quran-on-Chip
QR	Quick Response Code
Qu Hex	Quran using Hexadecimal
SME	Subject Matter Expert
SVD	Singular-Value Decomposition
UTF	Unicode Transformation Format
XOR	Exclusive OR