

Regeneration and Reuse of Content in the World Wide Web: a Model-based Approach

by

Edmund Balnaves

BA (University of Canberra) 1981

MBA (NTU) 1992

A thesis submitted in satisfaction of the requirements for the degree of Doctor of
Philosophy in Science

in the

School of Information Technologies

UNIVERSITY OF SYDNEY

Systematic Content Reuse in the World Wide Web: a Model-based Approach

May 2004

Apart from quotations from other works, all of which are duly acknowledged in the text, this thesis is entirely my own work.

Abstract

Content reuse is an important area of software engineering and information systems research. Key factors affecting systematic content reuse are the heterogeneous nature of digital encoded content; the need for ontological structures to coherently organise content for reuse, the need for workflow systems to facilitate content authoring, and the reality of continuous transformation that is demanded by regeneration and reuse of content. It is argued that the privileged focus on content reuse differentiates the Content Management System (CMS) from other information management systems, but that CMS products currently lack a systematic approach to the problem of content regeneration and reuse.

This thesis develops a conceptual model directed by a set of principles to facilitate systematic content reuse addressing the underlying complexities of heterogeneous media and presentational formats. It is argued that the axiom of separation of content, structure and presentational form must be realised not only in a document ontology but also in an information system architecture. A systematic approach to regeneration and reuse of digital multimedia requires a clear and comprehensive strategy on content composition, publishing runtime engine, information retrieval, and database management for loosely typed and unstructured data. This model is presented in four layers, each layer building in complexity on the lower layer, and each layer addressing specific aspects of the content regeneration and reuse problem. The Content Model for Reuse is explored through Software Development Research and a complex field trial. The results of a substantial field trial conducted over a three-year period are presented. The completeness of the model is evaluated against qualitative measures at each layer of the conceptual model. Measures for information currency and content reuse are defined and applied in the context of the field trial with results that indicate both better content currency over time and increased content reuse.

Acknowledgements

This thesis is inspired by and would not be possible without the stimulating and engaging intellectual environment of the University of Sydney and its staff. Special thanks go to Rob Gamsey of the Australian Broadcasting Corporation, who got me started on Content Management in the Wallace project. Special thanks also to my supervisor, Professor Jon Patrick, whose creative engagement with the world is daunting, and to Associate Professor Joseph Davis and the Knowledge Management Research Group. Finally, and most importantly, the British Council Australia provided a project that had endless possibilities, was deep in scope and inspired in design - and especially to Glenn Dudley and Gill Edwards, the original architects of the Montage concept.

TABLE OF CONTENTS

LIST OF FIGURES	X
LIST OF TABLES	XII
CHAPTER 1 – INTRODUCTION	1
1.1 BACKGROUND	1
1.2 RESEARCH QUESTIONS	9
1.3 SIGNIFICANCE OF THE STUDY AND TARGET AUDIENCE	10
1.4 INTELLECTUAL PROPERTY	11
1.5 THESIS ORGANISATION	12
CHAPTER 2 - LITERATURE REVIEW	13
2.1 DEFINITION OF TERMS	14
2.2 CONTENT REUSE AND SOFTWARE REUSE	19
2.3 DIGITAL DOCUMENT ARCHITECTURE	25
2.3.1 Encoding (text and multimedia)	26
2.3.2 Text-based Electronic Publishing	29
2.3.3 Multimedia Document Ontologies	35
2.4 CONTENT DISCOVERY AND TRANSFORMATION	37
2.4.1 Metadata Standards	37
2.4.2 Content Discovery	42
2.4.3 Content Sharing and Syndication	46
2.4.4 Content Transformation	47
2.4.5 Content Reuse = Multiple Ontologies	48
2.5 THE SYSTEMS CONTEXT	51
2.5.1 Digital Library Systems	52
2.5.2 Content Management Systems	52
2.5.3 Knowledge Management Systems	55
2.5.4 Document Management Systems	57
2.5.5 Learning Management Systems	58
2.5.6 Comparative Analysis Of Systems	60
2.6 CURRENT INDUSTRY PRACTICE	63
2.6.1 Context of the Review	63
2.6.2 CMS product classes	63
2.6.3 Industry reviews and reports	65

2.6.4 Systems Reviewed	67
2.6.5 Category 1: Proprietary > \$100,000	69
2.6.6 Category 2: Proprietary < \$100,000	72
2.6.7 Category 3: Open Source	79
2.6.8 Industry Practice and Content Reuse	83
2.7 THE <i>ECONOMICS</i> OF CONTENT REGENERATION AND REUSE	86
2.8 THE <i>MANAGEMENT</i> OF CONTENT REGENERATION AND REUSE	89
2.8.1 Content Modelling	91
2.8.2 Content Identification, Persistence and Discovery	92
2.8.3 Workflow Control of the Publishing Process	97
2.8.4 Summary	101
2.9 THE CHALLENGE	102
CHAPTER 3 - RESEARCH DESIGN AND METHODOLOGY	107
3.1 RESEARCH METHODOLOGY	107
3.1.1 Methods of Inquiry	107
3.1.2 Research Approach	111
3.2 EVALUATION	112
3.2.1 Software Development Research	112
CHAPTER 4- THEORETICAL MODEL	119
4.1 RELEVANCE OF THE MODEL	120
4.2 THE CONTENT MODEL FOR REUSE (CMR)	124
4.3 WEB SERVICES AS A INSTANCE OF THE CMR	127
4.4 LAYER 1 -THE ECONOMIC CONTENT UNIT	129
4.4.1 Definition	131
4.4.2 Interfaces	132
4.4.3 Discussion	134
4.5 LAYER 2 - DOCUMENT ONTOLOGY	137
4.5.1 Definition	138
4.5.2 Interfaces	139
4.5.3 Discussion	140
4.6 LAYER 3 -WORKFLOW MANAGEMENT AND AUTHORING	143
4.6.1 Definition	144
4.6.2 Interfaces	144
4.6.3 Discussion	150

4.7 LAYER 4 - TRANSFORMATION AND DELIVERY	156
4.7.1 Definition	159
4.7.2 Interfaces	160
4.7.3 Discussion	161
4.8 ALTERNATIVES TO THE MODEL	164
4.8.1 XML is not sufficient for Reuse	164
4.8.2 Model-driven content management	166
4.9 SUMMARY	170
CHAPTER 5 – PROTOTYPE SPECIFICATION	171
5.1 INTER-PUBLISH BACKGROUND AND OBJECTIVES	171
5.1.1 Prototype objectives	171
5.1.2 Prototype chronology	172
5.2 CMR PROTOTYPE – ARCHITECTURE	174
5.2.1 Overview	174
5.2.2 Architecture	175
5.2.3 Architectural Limitations	182
5.2.4 Focus of the Field Implementation	183
5.3 FUNCTIONAL DECOMPOSITION FOR REUSE	184
5.3.1 Layer 1 Functional Decomposition – Economic Content Units	185
5.3.2 Layer 2 Functional Decomposition – Document Ontology	187
5.3.3 Layer 3 Functional Decomposition – Workflow management	190
5.3.4 Layer 4 Functional Decomposition - Transformation and Delivery	191
5.3.5 Data Flow	191
5.4 INTER-PUBLISH DATABASE DESIGN	192
5.4.1 Background to the design of the content repository	192
5.4.2 Layer 1 Database implementation - the ECU	196
5.4.3 Layer 2 Database Implementation - the Document Ontology	201
5.4.4 Layer 3 Database implementation - the Workflow process	202
5.4.5 Layer 4 Database implementation – Transformation and Delivery	205
5.5 INTER-PUBLISH USER INTERFACE DESIGN	205
5.5.1 Browser-based authoring interface version 1 (September 1999)	205
5.5.2 Browser-based authoring interface version 2 (December 2000)	208
5.5.3 Browser-based authoring interface version 3 (April 2001)	209
5.5.4 Abstract site modelling	212
5.5.5 Prototype version 4 (April 2002)	213

5.5.6 Prototype version 5 (January 2003)	214
5.5.7 Workflow Management Concepts	215
5.5.8 Authentication	217
5.5.9 Controlled Vocabulary and Metadata	217
5.5.10 Confusion with fragmented content	218
5.5.11 Text mining of the ECU for authoring support	221
5.5.12 WebDAV authoring interface	224
5.5.13 FTP Authoring Interface	225
5.5.14 Web Services authoring interface	226
5.6 CONTENT TRANSFORMATION AND DELIVERY	227
5.6.1 Content delivery in the Field Implementation	227
5.6.2 Browser-based navigation and publishing interface	231
5.6.3 Inter-Publish Transformation Language	231
5.6.4 Syndication	233
5.6.5 XML profiles for Personalisation	233
5.7 TESTING, QUALITY ASSURANCE AND DOCUMENTATION	233
5.8 CONCLUSION	234
CHAPTER 6 – THE MONTAGE SERVICE	237
6.1 THE MONTAGE PROJECT	237
6.2 MONTAGE REQUIREMENTS – PHASE 1	240
6.2.1 Project Scope – Montage Website	240
6.2.2 Project data requirements	241
6.3 MONTAGEWORLD REQUIREMENTS – PHASE TWO	242
6.3.1 Project Scope – Content Reuse	242
6.3.2 Publishing Server	243
6.3.3 Publishing Software	243
6.3.4 Design aspects of the MontageWorld website	244
6.3 CONTENT REUSE	246
6.4 MONTAGE PROJECT OUTCOMES	247
CHAPTER 7 – EVALUATION	251
7.1 QUALITATIVE EVALUATION OF REUSE AGAINST THE MODEL	252
7.1.1 Content Reuse Outcomes from the Field Trial	252
7.1.2 Evaluation against the CMR layer 1	255
7.1.3 Evaluation against the CMR layer 2	256

7.1.4 Evaluation against the CMR layer 3	258
7.1.5 Evaluation against the CMR layer 4	260
7.1.6 Completeness of the Prototype against the Content Model for Reuse	261
7.2 QUANTITATIVE EVALUATION OF REUSE ANALYSIS	261
7.2.1 Resources used in the Reuse Analysis	261
7.2.2 Website Content Profiles	262
7.2.3 Metrics measuring Content Currency	265
7.2.4 Further Measurement of Content Reuse	268
7.2.5 Focal points for content reuse	273
7.3 POST IMPLEMENTATION SURVEY	275
7.3.1 Responses relating to "Ease of Use"	276
7.3.2 Responses relating to Task utility	277
7.3.3 Responses in relation to Job utility (Job related questions)	277
7.3.4 Post Implementation Review Summary	278
7.4 CONCLUSIONS FROM THE EVALUATION	278
CHAPTER 8 - SUMMARY AND CONCLUSION	283
8.1 SUMMARY OF FINDINGS AND CONTRIBUTIONS	284
8.2 POTENTIAL IMPLICATIONS FOR IS RESEARCH AND PRACTICE	286
8.2.1 Authoring and WYSIWYG	286
8.2.2 Detecting content suitable for systematic reuse	289
8.2.3 Open Source Directions	290
8.2.4 Performance	290
8.2.5 Future directions in the prototype	292
8.3 CONCLUSIONS	296
GLOSSARY	299
BIBLIOGRAPHY	311
APPENDIX 1 – INDUSTRY REVIEW	
APPENDIX 2 - WEB SERVICES DEFINITION	
APPENDIX 3 - INTER-PUBLISH DOCUMENTATION	
APPENDIX 4 - USER SATISFACTION SURVEY	

List of Figures

FIGURE 2.1: A WEB PAGE FROM THE MONTAGE WEBSITE	20
FIGURE 2.2: A PRESENTATION TEMPLATE ABSTRACTED FROM FIGURE 2.1.....	20
FIGURE 2.3: A CONTENT FRAGMENT FROM FIGURE 2.1.	21
FIGURE 2.4: 3RDMLIL HOME PAGE – 3RDGEN (2003A)	69
FIGURE 2.5: 3RDGEN TECHNICAL ARCHITECTURE – 3RDGEN(2003B)	70
FIGURE 2.6. GRAVITYMAX TECHNICAL ARCHITECTURE - GRAVITYMAX (2003).....	73
FIGURE 2.7: OBTree “OBJECT-ORIENTED” ENGINE - OBTree TECHNOLOGIES INC, (2003)	75
FIGURE 2.8: MICROSOFT CMS 2002 TECHNICAL ARCHITECTURE - MICROSOFT (2003).....	77
FIGURE 2.9: REDDOT TECHNICAL ARCHITECTURE - REDDOT (2003)	78
FIGURE 2.10: COCOON ROADMAP	81
FIGURE 3.1: SYSTEMS DEVELOPMENT RESEARCH	110
FIGURE 4 1: CMR: A MODEL FOR CONTENT MANAGEMENT, REGENERATION AND REUSE	123
FIGURE 4.2: DEFINITION OF THE ECU.....	131
FIGURE 4.3: ECU SCHEMA.....	133
FIGURE 4.4: ECU SELECTION RULE.....	133
FIGURE 4.5: DOCUMENT ONTOLOGY	139
FIGURE 4.6: SCHEMA DEFINING THE GENERALISED DOCUMENT ONTOLOGY SPECIFICATION.....	141
FIGURE 4.7: CONTENT WORKFLOW FRAMEWORK	144
FIGURE 4.8: SCHEMA FOR CONTENT RETRIEVAL AND STATIC COLLECTION MANAGEMENT	148
FIGURE 4.9: THE DOCUMENT COLLECTION.....	149
FIGURE 4.10: DOCUMENT SELECTION RULE (FOR WEB SERVICES RETRIEVAL).....	149
FIGURE 4.11: SCHEMA RELATING TO WORKFLOW MANAGEMENT	149
FIGURE 4.12: CONTENT TRANSFORMATION AND DELIVERY	159
FIGURE 4.13: CONTENT TRANSFORMATION RULE-SETS	161
FIGURE 4.14: CONTENT DISTRIBUTION AND SYNDICATION SCHEMA	161
FIGURE 5.1: INTER-PUBLISH ARCHITECTURE.....	177
FIGURE 5.2: CMR WEB SERVICES XML SCHEMA.....	179
FIGURE 5.3: ILLUSTRATIVE DATA FLOW MODEL FOR MULTIMEDIA PUBLISHING	192
FIGURE 5.4: FIRST PROTOTYPE RELATIONAL MODEL.....	194
FIGURE 5.5: STRUCTURE OF THE ECU	197
FIGURE 5.6: ECONOMIC CONTENT UNIT SCHEMA.....	200
FIGURE 5.7: DOCUMENT ONTOLOGY SCHEMA	204
FIGURE 5.8: SYNDICATION SCHEMA	206

FIGURE 5.9: THREE-PANE CONTENT EDITING DESIGN.....	207
FIGURE 5.10: TWO-PANE CONTENT DISSECTION.....	209
FIGURE 5.11: TWO-WINDOW DESIGN – SITE NAVIGATION PANE AND CONTENT ACTIVITY PANE....	210
FIGURE 5.12: CONTEXT-SENSITIVE VISUAL EDITING IN A MULTI-LINGUAL ECU	215
FIGURE 5.13: WORKFLOW EXAMPLE.....	216
FIGURE 5.14: BAG OF WORDS	222
FIGURE 5.15: SIMILARITY MEASURE	222
FIGURE 5.16: JACQUARD COEFFICIENT.....	223
FIGURE 5.17: CONTENT REUSE IN E-ZINES.....	228
FIGURE 5.18: SIMPLE TRANSFORMATION EXAMPLE	232
FIGURE 5.19: INTERNAL ALERTS	234
FIGURE 6.1: HIERARCHICAL TEMPLATE STRUCTURES.....	246
FIGURE 7.2: WEBSITE CONTENT TURNOVER AND INFORMATION TURNOVER.....	267
FIGURE 7.3: ECU REUSE BY DOCUMENT TYPE	272
FIGURE 7.4: SUMMARY OF ECU REUSE.....	273
FIGURE 7.5: RESPONSES ON USER INTERACTION WITH INTER-PUBLISH.....	276
FIGURE 7.6: TASK UTILITY IN PUBLISHING.....	277
FIGURE 7.7: RESPONSES ON UTILITY IN RELATION TO JOB ROLE.....	278

List of Tables

TABLE 2.1: CONTRASTING SYSTEMS APPROACHES	61
TABLE 2.2: SYSTEMS REVIEWED	68
TABLE 2.3: PRINT CITATIONS VS. URL CITATIONS	93
TABLE 3.1: CONTRASTING ACADEMIC RESEARCH AND PRACTICE.....	107
TABLE 4.1: WEB SERVICES FUNCTIONS AT LAYER 1	133
TABLE 4.2: WEB SERVICES FUNCTIONS AT LAYER 2	140
TABLE 4.3: WEB SERVICES FUNCTIONS AT LAYER 3	147
TABLE 4.4: WORKFLOW EVALUATION MEASURES (PEREZ & ROJAS, 2000).....	155
TABLE 4.5: WEB SERVICES FUNCTIONS AT LAYER 4	160
TABLE 5.1: FUNCTIONAL MAPPING TO LAYER 1 OF THE CMR.....	186
TABLE 5.2: FUNCTIONAL MAPPING TO LAYER 2 OF THE CMR.....	188
TABLE 5.3: FUNCTIONAL MAPPING TO LAYER 3 OF THE CMR.....	189
TABLE 5.4: FUNCTIONAL MAPPING TO LAYER 4 OF THE CMR.....	190
TABLE 6.1 INITIAL MONTAGE DATA REQUIREMENTS.....	241
TABLE 6.2: COUNTRIES WITH PARTICIPATING SCHOOLS	248
TABLE 7.1: SUMMARY OF CONTENT REUSE AT EACH LAYER	254
TABLE 7.2: EVALUATION AGAINST THE CMR LAYER 1	256
TABLE 7.3: EVALUATION AGAINST THE CMR LAYER 2	258
TABLE 7.4: EVALUATION AGAINST THE CMR LAYER 3.....	259
TABLE 7.5: EVALUATION AGAINST THE CMR LAYER 4.....	260
TABLE 7.6: PUBLISHED WEBSITE CONTENT SUMMARY FOR MONTAGE SITES	263
TABLE 7.7: CONTENT AUTHORING PROFILE BY MONTH FOR 2001,2002 AND 2003.....	264
TABLE 7.8: NUMBER OF DOCUMENTS BY DOCUMENT TYPE, 2001 - 2003.....	270
TABLE 7.9: CONTENT REUSE BY DOCUMENT TYPE	272
TABLE 8.1: FUTURE DIRECTIONS (LAYER 1).....	293
TABLE 8.2: FUTURE DIRECTIONS (LAYER 2).....	294
TABLE 8.3: FUTURE DIRECTIONS (LAYER 3).....	295
TABLE 8.4: FUTURE DIRECTIONS (LAYER 4).....	296