Validation of the CIDOC CRM Using both Extended Graphical and Category Theory Representations: Includes two New Zealand Case Studies

A Thesis presented in partial fulfilment of the requirements for the degree of

Master of Business Studies

in

Information Systems

at Massey University, Palmerston North, New Zealand

Jia Zhou

2004

Abstract

There is considerable interest in the use of the Internet to retrieve and integrate multimedia information from centres of cultural heritage such as museums and art galleries. The ultimate desire of most devotees of cultural matters is to have universal access, through a single portal, to detailed information from sites throughout the world. This level of interoperability is not an easy task both technically and culturally. To provide an avenue where some of the technical problems of accessing information from a huge range of unique database environments can be resolved, a semantic conceptual reference model (CRM) was proposed by The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC). The model provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. It is intended to provide a common and extensible semantic framework to which any cultural heritage information can be mapped. In this research two methods are proposed and developed to support the validation of the Conceptual Reference Model. The methodologies, one graphical and the other based on category theory, are used to replicate three published international validation activities and two new validations based on information supplied by two New Zealand heritage sites. This report also includes a literature review describing the main ideas and structures that form the basis of the CRM.

Acknowledgements

This has been an interesting but complex research activity. I would like to thank Judith Taylor of the Suter Galley in Nelson, and Susanne Geiser and Lynne Trafford of the Te Manawa Museum and Science Centre for their support and supply of records of cultural artefacts. I would also like to thank Bill Vernon and Abby Turbott of Vernon System for their assistance not only with this thesis, but research carried out at the honours level. I would like to thank Matthew Jackson for his support in applying Category Theory in this thesis.

I must not forget the support provided by the Information Systems Department and its teaching, technical and general staff.

I give my thanks to my friend Judy Engelbrecht for her support and friendship, my husband for his help on the format of this thesis and my family for their support and understanding during thesis writing.

Finally, I would like to express my thanks to my supervisor Barry Jackson for his support, encouragement and patience during the whole process of my research thesis.

Table of Contents

Page

Abstract i
Acknowledgements ii
Table of Contents iii
List of Figures vii
List of Tables vii

CHAPTERS:

1 Introduction 1
1.1 Background 1
1.2 Importance 1
1.3 Issues 2
1.4 Research problem 2
1.5 Aims and objectives 3
1.6 Research process 3
1.7 Report structure 4

2 Research Design 7
2.1 Introduction 7
2.2 Methodology 7
2.2.1 Research the field of study 7
2.2.2 Develop a model or framework (In this case a graphical tool and mathematical documentation mapping tool) 8
2.2.3 Test the model (in this case the CRM and the graphical and mathematical filter) 8
2.2.4 Evaluate and reflect 9

3 Literature Review 10
3.1 Introduction 10
3.2 Needs of Interoperability of Cultural Heritage Information 10
3.2.1 Introduction 10
3.2.2 Challenge of information integration 11
3.2.3 Need for semantic interoperability 12
3.3 Foundation of CRM: the Use of Ontology 14
3.3.1 Introduction 14
3.3.2 Ontology 14
3.3.3 Uses of Ontology in Information Systems 17
3.3.4 Features of CRM Ontology 18
3.3.4.1 Integration of Context-Free propositions 19
3.3.4.2 Monotonicity (Extendibility) 20
3.4 The CIDOC Conceptual Reference Model 23
3.4.1 Who is involved? 23
3.4.2 Introduction of CIDOC CRM 23
3.4.2.1 Overview of CRM Structure 24
3.4.2.2 The CRM Design Process 37
3.4.2.3 Versions of CRM 38

4 Analysis 1: Diagrammatical and Notational Representation 40
4.1 Introduction 40
4.2 Diagrammatical representation 40
4.2.1 Introduction 40
4.2.2 Definition 40
4.2.3 Taxonomy 41
4.2.4 Example 45
4.2.5 Summary 47
4.3 Mathematical representation 48
4.3.1 Introduction 48
4.3.2 Definition, Basic Concept and Syntax 48
4.3.3 Examples 49
4.3.4 Summary 56

5 Analysis 2: International Validation of the CRM 58
5.1 Mapping of the generic elements in EAD to CRM (based on CRM Version 3.0) 58
5.1.1 Introduction to EAD 59
5.1.1.1 Structure of <eadheader> in EAD: 59
5.1.1.2 Structure of <frontmatter> in EAD: 61
5.1.1.3 Structure of < archdesc > in EAD: 61
5.1.2 Mapping Formalism 63
5.1.2.1 Differences between EAD DTD and CRM 63
5.1.2.2 Mapping scheme 64
5.1.3 Mapping of the Encoded Archival Description (<ead>) tag 66
5.1.4 Mapping of the elements of <eadheader> 66
5.1.4.1 Mapping of <eadheader>: 66
5.1.4.2 Mapping of <eadid>: 68
5.1.4.3 Mapping of <titleproper>, <subtitle>: 69
5.1.4.4 Mapping of <date>: 69
5.1.4.5 Mapping of <author>: 71
5.1.4.6 Mapping of <publisher>: 72
5.1.4.7 Mapping of <address>: 72
5.1.4.8 Mapping of <publicationstmt>: 73
5.1.4.9 Mapping of <publisher>, <date>, <address> and <num>: 73
5.1.4.10 Mapping of <note> and <editionstmt>: 74
5.1.4.11 Mapping of <seriesstmt>, <titleproper>, <num> and <p>: 75
5.1.4.12 Mapping of <langusage>, <language>: 76
5.1.4.13 Mapping of <creation>: 77
5.1.4.14 Mapping of <revisiondesc>, <item>, and <list>: 78
5.1.5 Mapping of the elements of <frontmatter>: 78
5.1.6 Mapping of the elements of <archdesc>: 79
5.1.6.1 Mapping of <did>.<head> and <did>.<unittitle>: 79
5.1.6.2 Mapping of <did>.<unitdate>: 80
5.1.6.3 Mapping of <did>.<unitid>: 80
5.1.6.4 Mapping of <did>.<origination>: 81
5.1.6.5 Mapping of <did>.<physdesc>: 81
5.1.6.6 Mapping of <did>.<repository> and <did>.<physloc>: 82
5.1.6.7 Mapping of <did>.<abstract> and <did>.<note>: 83
5.1.6.8 Mapping of <controlaccess> elements 83
5.1.6.9 Mapping of <dscgrp> elements 83

5.2 Mapping of the Dublin Core Element set to the CIDOC CRM (based on Dublin Core Version 1.1 and CRM Version 3.0) 84
5.2.1 Introduction to Dublin Core 84
5.2.2 Mapping Formalism 87
5.2.2.1 Comparison between DC and CRM 87
5.2.2.2 Mapping scheme 88
5.2.3 Mapping Dublin Core Metadata Element Set to CRM 89
5.2.3.1 Mapping of the Dublin Core described resource. 90
5.2.3.2 Dublin Core Qualifiers 91
5.2.3.3 Mapping of DC.Title 92
5.2.3.4 Mapping of DC.Agents 93
5.2.3.5 Mapping of DC.Creator and DC.Contributor 94
5.2.3.6 Mapping of DC.Subject 95
5.2.3.7 Mapping of DC.Description 97
5.2.3.8 Mapping of DC.Publisher 98
5.2.3.9 Mapping of DC.Date 99
5.2.3.10 Mapping of DC.Type 102
5.2.3.11 Mapping of DC.Format 103
5.2.3.12 Mapping of DC.Identifier 104
5.2.3.13 Mapping of DC.Source 105
5.2.3.14 Mapping of DC.Language 105
5.2.3.15 Mapping of DC.Relation 106
5.2.3.16 Mapping of DC.Right 112
5.3 Mapping of the AMICO data model to the CIDOC CRM version 2.3 112
5.3.1 Introduction to AMICO 112
5.3.1.1 Data dictionary format 113
5.3.1.2 Category record 113
5.3.1.3 Associated multimedia files 114
5.3.1.4 Metadata record 114
5.3.2 Mapping formalism 114
5.3.2.1 Mapping scheme 115
5.3.3 Mapping AMICO data to CRM 115
5.3.3.1 Mapping for "Unique Identification" 115
5.3.3.2 Mapping for "What is it?" 115
5.3.3.3 What is it called? 117
5.3.3.4 What does it look like? 118
5.3.3.5 Who made it? 122
5.3.3.6 When was it made? 125
5.3.3.7 Where was it made? 126
5.3.3.8 What is it about? 126
5.3.3.9 What does it mean? 130
5.3.3.10 Who showed it? 130
5.3.3.11 Who owned it? 130
5.3.3.12 What is it related to? 133
5.3.3.13 Who documented it? 136
5.4 Summary 137

6 Analysis 3: New Zealand Validation of the CRM (Two New Zealand Studies) 138
6.1 Introduction 138
6.2 Mapping of Suter Gallery Data Model to CRM Version 3.0 138
6.2.1 Introduction of Suter Gallery collection and Data Model 138
6.2.2 Mapping formalism 141

6.2.3 Mapping Suter Gallery Data to CRM 141
6.2.3.1 Mapping for Unique identification 141
6.2.3.2 Mapping for "What is it?" 141
6.2.3.3 Mapping for "What is it called?" 142
6.2.3.4 Mapping for "What does it look like?" 142
6.2.3.5 Mapping for "When was it made?" 146
6.2.3.6 Mapping for "Who owned it?" 147
6.2.3.7 Mapping for "What is it about?" 149
6.2.3.8 Mapping for "Who made it?" 150
6.2.3.9 Mapping for "What is it related to?" 151
6.2.3.10 Mapping for "Who documented it?" 152
6.2.3.11 Mapping for "Multimedia Metadata Fields" 153
6.3 Mapping of Te Manawa Data Model to CRM Version 3.0 154
6.3.1 Introduction of Te Manawa Collection and Data Model 154
6.3.2 Mapping Formalism 157
6.3.3 Mapping Te Manawa Data to CRM 157
6.3.3.1 Mapping for Unique Identification 158
6.3.3.2 Mapping for "Object" 158
6.3.3.3 Person 168
6.3.3.4 Documentation 169
6.3.3.5 Photo/Audio-Visual 170
6.3.3.6 Event 170
6.3.3.7 Other Authority Files 171
6.4 Summary 174

7 Conclusions 175
7.1 Aims 175
7.2 Future Research 179

8 References 180

9 Appendices 185
9.1 Appendix 1 185
9.1.1 The Entity List (Selection) 185
9.1.2 E5 Event 186
9.1.2.1 Properties 186
9.1.2.2 The entity is referenced by: 187
9.1.2.3 The entity inherits references: 187
9.1.3 E19 Physical Object 187
9.1.3.1 Properties: 187
9.1.3.2 The entity is referenced by: 189
9.1.3.3 The entity inherits references: 189
9.1.4 E7 Activity 189
9.1.4.1 Properties 190
9.1.4.2 The entity is only referenced by itself 191
9.1.4.3 The entity inherits references: 191
9.2 Appendix 2 192
9.3 Appendix 3 195
9.4 Appendix 4 201
9.5 Appendix 5 206

List of Figures

Figure 1 Research Process 4
Figure 2 Comparison of accessing various database structures 14
Figure 4 Good Minimal Ontological Commitment (Guarino, 2003) 16
Figure 5 Richer System and Poorer System – based on Doerr and Crofts (1999) 22
Figure 6 CRM Main Entities based on CRM Version -- based on Doerr and Crofts (1999) 25
Figure 7 CRM Struture (Version 3.3.2) 28
Figure 8 Relationships of some CRM Entities -- based on Doerr and Stead (2002) 30
Figure 9 Relationship of Entity Condition Assessment -- Based on Doerr and Crofts (1998) 32
Figure 10 Mapping of "Treaty of Waitangi" to CRM 33
Figure 11 Actual Path of connections of 'Activity' to 'Image' and 'Creation' to 'Document' 35
Figure 12 Hierarchical Structure of Related Entities in the mapping of "Treaty of Waitangi" .. 36
Figure 13 Representation-related taxonomic aspects (Blackwell & Engelhardt, 2004) 42
Figure 14 Contextual taxonomic aspects (Blackwell & Engelhardt, 2004) 43
Figure 15 Example of non-physical relationship connects by inherited link 47
Figure 16 Hierarchy Structure of <eadheader> 60
Figure 17 Hierarchy Structure of <archedesc> 62

List of Tables

Table 1 Nine aspects of diagrams and diagram use (Blackwell & Engelhardt, 1998) 41
Table 2 The DC Element (DC Metadata Initiative (2003) 85
Table 3 DCMI Type Vocabulary ([DCT1]) and their corresponding CRM entities 91
Table 4 Suter Gallery Data Fields 139
Table 5 Te Manawa Museum Data Fields 155

1 Introduction

1.1 Background

The World Wide Web has transformed the way information objects are distributed and presented. Generally speaking centres of cultural heritage, primarily museums and galleries, have been quick to take advantage of the new technology, even to the extent that many of them manage their own web sites. Most of these sites tend to be simply showcases providing little more than links to a number of images, yet it is well known that what is displayed is only a very limited view of what is stored within collections. Information remains hidden from the public view, not because of a lack of enthusiasm on the part of the centres but because of the perceived difficulties in harvesting the information in a coherent and informative manner.

Some museums and art galleries have begun to investigate ways of exploiting the web as a global resource of cultural heritage information. Some have chosen to adopt an agreed metadata format (e.g. Dublin Core), while others have sought to create a 'universal' data model against which existing databases could be mapped. However, whatever approach is adopted, museums will have to establish solid and reliable systems that support the integration and distribution of rich and varied information contained in their collection systems.

1.2 Importance

Many observers believe that it is a mark of a civilised society to have access to objects of cultural interest and value. Cultural artefacts are clearly not located in one museum, or even one country or continent; they are dispersed in a variety of environments around the world. Gaining access and integrating associated information from such complex and dispersed environments requires electronic interoperability. The current situation regarding interoperability is one of uncertainty with a number of concepts and ideas under investigation by numerous research organisations throughout the world. Given the broad spectrum of approaches and ideas currently being researched, focussing on one contemporary approach could be seen as sensible and practical. It is with this pragmatic view in mind that the scope of thesis is confined to the object oriented semantic modelling approach pioneered by the CIDOC (International

Documentation Committee), a research group of the International Council of Museums (ICOM).

1.3 Issues

The major issues facing interoperability between centres of cultural heritage is the semantic and structural incompatibility of existing systems. Internationally based institutions have organised and structured their data in a number of different and sometimes unique ways. Even with the same collection management system (CMS), museums may have chosen to name and arrange entities and fields in quite different ways. Given the interests and values of individual archivists, it is quite likely that different levels of detail have been used to describe their collection. Even if the structures are compatible, terminology is often incompatible. The majority of solutions to this problem of incompatibility have been based on local transformation rules, or have adopted minimalist systems consisting of core data, and as a result have lost much of the richness contained in the original information.

For potential users of the CIDOC CRM the apparent complexity of the model and how this model might be mapped to existing data structures are important issues. These issues are addressed, in part, within this thesis.

1.4 Research problem

The research in this thesis centres on the object oriented Conceptual Reference Model developed by the document standards group of the ICOM/CIDOC. The aim of the CRM is to provide ways for museums to render their information resources to one another without losing detail or precision.

The problem faced by this researcher and many archivists and information systems professional wishing to understand and work in this field is the complexity and utility of the CRM framework.

It is intended in this research to address this problem in four ways:

- Undertake a literature review bounded by the CRM perspective
- Develop tools to enable researchers map archival data to the CRM
- Use the new tools to replicate the CRM validation exercises published in the international research literature

- Apply the new tools to two New Zealand centres of cultural heritage.

1.5 Aims and objectives

The aims of the research are:

1. To gain an understanding of the CIDOC Conceptual Reference Model (CRM) used to represent the semantic content of cultural data held within museums and art galleries.

   Note: The scope of this objective is primarily limited to published research undertaken by The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) and the Institute of Computer Science of the Foundation for Research and Technology – Hellas (ICS-FORTH) under the direction of Martin Doerr.

2. To develop new tools to assist in the validation of the CRM against real-world heritage collection systems.

   Note: An extension to the graphical tool used by Martin Doerr is proposed as well as a new mathematical tool based on category theory notation.

3. To investigate, using the above mentioned graphical and mathematical notation, several of the seminal publications used to validate the CRM on an international level.

   Note: The international cultural heritage systems being; Encoded Archival Description (EAD), Dublin Core (DC) and Art Museum Image Consortium (AMICO).

4. To apply the same graphical and mathematical notation to validate the CIDOC CRM within the New Zealand context.

   Note: The two New Zealand centres of cultural heritage being; The Suter Gallery in Nelson and Te Manawa, a Museum and Science Centre in Palmerston North.

1.6 Research process

In general terms the process adopted in this research follows that proposed by Bourner (1996, p7). There are four steps in Bournier's approach:

1 Research the field of study

2 Develop a model or framework

3 *Test the model*

4 *Undertake evaluation and reflection*

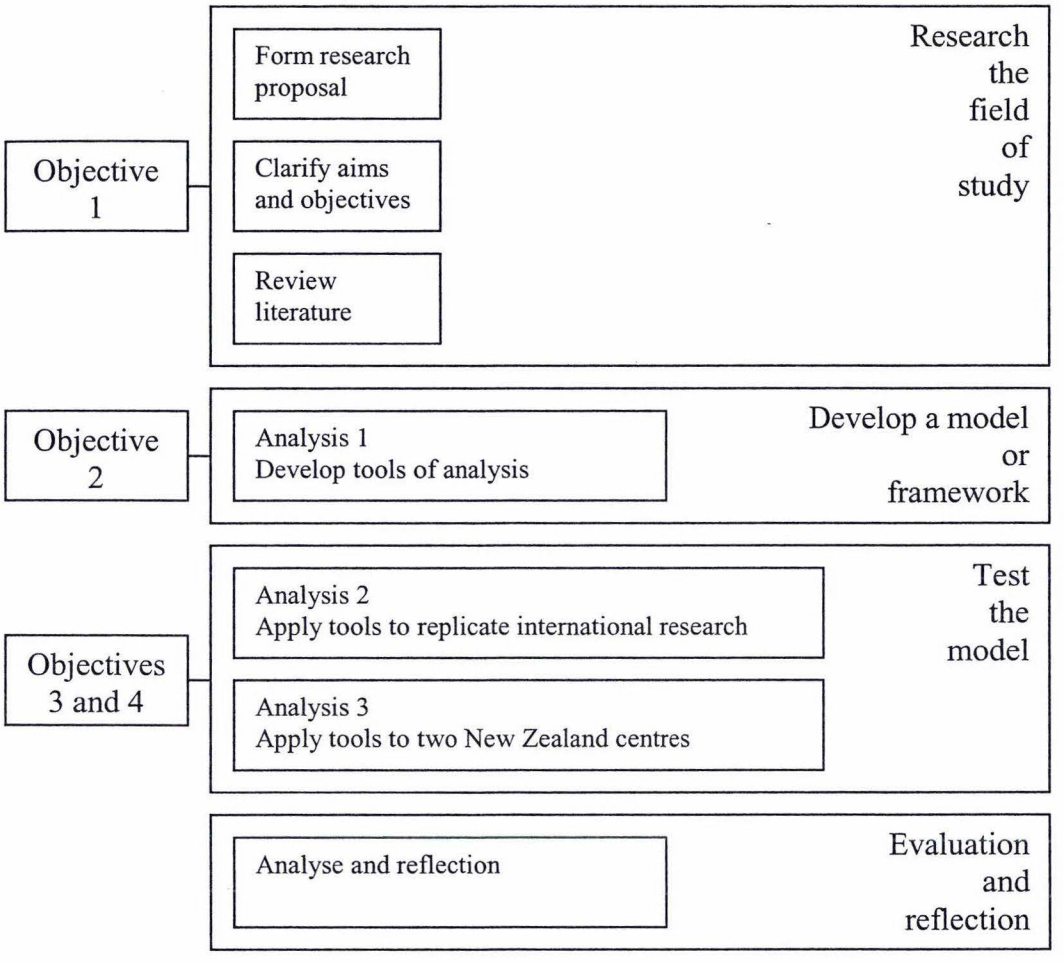

Figure 1 Research Process

1.7 Report structure

The report has the following structure, which is aligned with the Aims and Objectives set out in Section 1.5 and the Research Process in Section 1.6.

Chapter 1: Introduction

The rationale for undertaking this research is outlined in this chapter, together with a description of the research problem and lists the main aims of the research.

Chapter 2: Research design: Methodology

The 'four step research approach' proposed by Bournier provides the guiding framework for the research design. In fact, the research processes are aligned not only to the research process but also to the structure of this report.

Bournier's steps		Research Report	
		Chapter Numbers	Chapter topic
1	Research the field of study	3	Literature Review
2	Develop a model or framework	4	Validation Tools
3	Test the model	5	International validation
		6	New Zealand validation
4	Undertake evaluation and Reflection	7	Conclusions

Chapter 3: Literature Review (Introduction)

The chapter is concerned with discussing the origin of the CIDOC Conceptual Reference Model, in particular why such a model is needed to facilitate the sharing of cultural information.

Chapter 4: Analysis 1: Validation tools -- Graphical and Mathematical Notation

The need to view the validation exercises in a consistent manner was the motivation to develop both graphical and mathematical representations.

Note: There is a possibility that applying category theory to the CRM domain could lead to an effective parsing algorithm to link cultural databases to the CRM. Such an opportunity is outside the scope of this research.

Chapter 5: Analysis 2: International Validation of the CRM

This chapter brings to bear the graphical and mathematical notation, developed in Chapter 4, to explore and replicate three representative investigations undertaken by researchers to validate the CRM.

Chapter 6: Analysis 3: New Zealand Validation of the CRM

Essentially this chapter applies the techniques used in chapter 5 to two New Zealand centres of cultural heritage.

Chapter 7: Discussions, Conclusions and Recommendations

The final chapter provides a summary of the findings of this research and discusses their relevance to the main aims stated in Chapter 1.

Bibliography

All the references emerged in the report are listed in this section.

Appendices

The appendices contain information from the CIDOC CRM. The data are only a small sample of what is provided in the original CRM documentation. It provides reference documentation to support the validation processes described in this thesis. Care needs to be exercised as the CIDOC CRM has been modified several times in the recent past and different versions are referenced in different sections of this report.