Metadata Quality in
the Cultural Heritage Sector :
Stakes, Problems and Solutions

Thèse présentée en vue de l'obtention
du grade de Docteur en
Sciences de l'Information et de la Communication

Directeur de thèse :
Isabelle BOYDENS

Seth VAN HOOLAND
Année académique 2008–2009
Contents

1 Introduction 1
  1.1 General introduction .............................................. 2
     1.1.1 Premise defended within the dissertation ............... 2
     1.1.2 Relevance of the research .................................. 3
     1.1.3 Definition of the key concepts ............................... 4
  1.2 Theoretical framework ........................................... 8
     1.2.1 Introducing the stratified time concept ............... 10
     1.2.2 Applying the stratified time to a concrete example .... 11
     1.2.3 Interaction between the different time levels .......... 17
  1.3 Working method .................................................. 18
     1.3.1 Literature overview ......................................... 19
     1.3.2 Case studies ................................................ 26
  1.4 Structure of the dissertation ................................... 32
  1.5 Original findings and outcomes ................................ 36

I Understanding metadata quality: a critical and original state of the art 39

2 Specificity of digitized cultural heritage 43
  2.1 Introduction ..................................................... 44
  2.2 Unclear identification of the users and their needs .......... 44
     2.2.1 Existing research on needs assessment .................. 45
     2.2.2 Practical problems with user needs assessment ........... 47
     2.2.3 Conceptual problems with user needs assessment ....... 48
  2.3 Role of cultural heritage within society ....................... 53
     2.3.1 Increasing the social relevancy of heritage by digitization ............... 54
     2.3.2 Increasing the economical relevancy of heritage by digitization ............... 55
  2.4 Dealing with connotative content ................................ 58
     2.4.1 Central role of images within digitization projects ....... 61
     2.4.2 Specificity of images as documentation aid ............ 63
  2.5 Conclusions ...................................................... 64
## CONTENTS

### 3 Applying ICT for metadata creation: an overview

3.1 Introduction ................................................. 69
3.2 Mismatched expectations ................................. 70
   3.2.1 Overview of the evolutions .......................... 70
   3.2.2 Confronting expectations with reality ............. 74
3.3 Metadata: old wine in new bags? ....................... 76
   3.3.1 Overview of the evolution of metadata ............... 77
   3.3.2 Specific metadata schemes for the cultural heritage domain 81
3.4 Impact of ICT on metadata and metadata practitioners .... 84
   3.4.1 Consequences of ICT on metadata ................... 84
   3.4.2 Professional status of metadata practitioners ........ 87
3.5 Current developments ................................. 88
   3.5.1 Ontologies ........................................ 89
   3.5.2 Folksonomies .................................... 93
   3.5.3 Use-neutral metadata and mash-ups .................. 96
3.6 Conclusions .............................................. 98

### 4 Trends in metadata quality research

4.1 Introduction .............................................. 101
4.2 Metadata quality in other application domains ............ 102
4.3 Quality control in library cataloging ...................... 106
   4.3.1 Development of copy cataloging .................. 106
   4.3.2 Defining quality in terms of speed and costs of access 107
   4.3.3 Google Books project .............................. 108
4.4 First steps outside the library community .................. 112
4.5 Good enough approach: accepting bad quality metadata .... 116
4.6 Conclusions .............................................. 117

### II Dealing with metadata quality in practice: three original strategies

#### 5 Applying data-profiling upon metadata records

5.1 Introduction .............................................. 123
5.2 Background of data-profiling in other domains ............ 124
5.3 Data-profiling within the cultural heritage sector .......... 126
5.4 Case study: the Royal Museum of Central Africa .......... 127
   5.4.1 Presentation of the analyzed metadata set and its context 128
   5.4.2 Presentation of the data-profiler tool ................ 131
   5.4.3 Results of the analyses ............................ 131
5.5 Conclusions .............................................. 138

#### 6 Correcting metadata with user comments

6.1 Introduction .............................................. 141
6.2 Possible strategies toward metadata correction ............... 142
6.3 Cost-constrained metadata quality enhancement ............. 144
Acknowledgements

First and foremost, I would like to thank Isabelle Boydens, the director of this dissertation. Her groundbreaking monography “Normes, informatique et temps” has been a major source of inspiration for me and I am truly grateful for the opportunity she has given me to build upon her work. As my PhD advisor, she also encouraged me to continuously improve my work through her rigorous comments upon the content and the structure of this dissertation. My thanks also go out to Françoise D’Hautcourt, head of the Information and Communication Science department of the ULB, for supporting my work throughout the years.

From the very start of this dissertation, I considered the interaction with metadata practitioners as a necessity to deliver original and operational outcomes. The time I have spent with catalogers, database managers, private consultants, software developers, digitization experts and policy makers from different countries has resulted in the empirical and operational part of this dissertation. My thanks go out to all of these metadata professionals who allowed me to gain a better understanding of the real-life constraints with the implementation of technologies for metadata creation and management in the cultural heritage sector. In particular, I would like to express my gratitude to a few professionals. Seth Kaufman, director of Whirl-i-Gig and head of the development team of the open source project CollectiveAccess, delivered an extremely valuable input. He programmed and implemented the interactive search interface prototype, presented in Chapter 7, and facilitated the contact with metadata practitioners in the United States. He holds a rare combination of wit, pragmatism and genuine interest in providing a better access to scientific and cultural heritage, all characteristics which make him one of the best ICT professionals I have ever met. Closer to home, I would like to thank Hein Vanhee, head of Collection Management of the Royal Museum for Central Africa. His expertise with the collection registration database of ethnographical objects of the museum formed the basis of presentations, publications and one of the case studies of the dissertation. The numerous discussions we had regarding metadata and cultural heritage management in general provided valuable input. I also greatly appreciated the feedback and encouragement of Saskia Scheltjens, head of the Arts and Humanities Library of the University of Ghent. Her extensive knowledge of both the technical and conceptual side of metadata
management in libraries helped me define my own approach to metadata.

On an academic level, my work has greatly benefited from the help of Eva Méndez, head of the Master in Digital Information of the Universidad Carlos III de Madrid, and Jane Greenberg, head of the Metadata Research Center at the University of North Carolina at Chapel Hill. Ever since our first encounter while reviewing European project proposals, Eva Méndez has continuously provided comments on my work and introduced me to relevant literature and other researchers. Through her I met Jane Greenberg, who invited me in 2007 to the School of Information and Library Studies at Chapel Hill. I treasure the encounters and discussions I had with Jane Greenberg and other researchers at Chapel Hill among the most valuable moments of my work as a doctoral student.

I would also wish to thank Karen Torres for proofreading my dissertation and her hospitality in New York.

Finally I would like to thank Mauve Carbonell, Pieter Vanslembrouck, Ann Clé, everyone from Fiersstraat 48 and Jo Seurync for their friendship and support, providing me with the energy and the motivation to focus on the completion of this dissertation. Thank you all.
List of Figures

1.1 Scan of three colored glassplate negatives made from a Rembrandt painting ........................................ 7
1.2 Scheme representing the interdependencies between digitization goals, resources, and processes. ................. 9
1.3 Overview of the different cataloging tools used throughout the history of the city museum of Tienen, part 1 .......... 14
1.4 Overview of the different cataloging tools used throughout the history of the city museum of Tienen, part 2 .......... 15
1.5 Illustration of the Query by Image Content search interface, as implemented in the search interface of the Hermitage state museum ........................................... 23
1.6 Overview of the different strategies for the selection of samples and cases, as defined by Bent Flyvbjerg ............... 29

2.1 Screenshot of the webshop of the art image site Lukasweb, which distributed high-quality digital reproductions of Flemish art .................................................. 49
2.2 Screenshot illustrating user comments sent to the local community-driven website “Heuvelland Verbeeldt” .............. 55
2.3 Online portfolio of the photographers of the Magnum cooperative photo agency ............................................ 57
2.4 Illustration of the different abstract and emotional terms Magnum has created within its thesaurus ..................... 58
2.5 Illustration of the use of racial origin descriptors (“type humain blanc”) within the indexing process at Magnum ........ 59
2.6 Panzani commercial used by Roland Barthes to illustrate the difference between denotation and connotation ........ 60
2.7 Representation of the “Black Square” from Kazimir Malevich (1913) which highlights the importance of connotative content of cultural heritage ................................................. 61

3.1 Hype cycle for metadata management as published by Gartner .......................................................... 79
3.2 Collapsed view of a METS record, illustrating the concept of container metadata standards .......................... 82
3.3 Modeling acquisition information based upon CIDOC CRM ....................................................................... 83
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Screenshot of the collection registration database of AMSAB, with a majority of empty metadata fields</td>
<td>86</td>
</tr>
<tr>
<td>3.5</td>
<td>Search- and display interface of the catalog of the public Flemish libraries</td>
<td>91</td>
</tr>
<tr>
<td>3.6</td>
<td>View of the wordcloud</td>
<td>92</td>
</tr>
<tr>
<td>3.7</td>
<td>View of the authority file concerning author names</td>
<td>92</td>
</tr>
<tr>
<td>3.8</td>
<td>View of the thesaurus</td>
<td>92</td>
</tr>
<tr>
<td>3.9</td>
<td>Results of the automated matching of terms between local and internationally used thesauri</td>
<td>93</td>
</tr>
<tr>
<td>3.10</td>
<td>Screenshot of the tagging interface of the Steve-project</td>
<td>94</td>
</tr>
<tr>
<td>3.11</td>
<td>Screenshot of the tagcloud of the Archives de Normandie project, representing the most popular tags</td>
<td>95</td>
</tr>
<tr>
<td>3.12</td>
<td>Screenshot of the development plan of the Coney Island site in New York, which confronts historical images with the actual state of the site</td>
<td>99</td>
</tr>
<tr>
<td>4.1</td>
<td>Data cleaning tool from the biodiversity domain</td>
<td>103</td>
</tr>
<tr>
<td>4.2</td>
<td>Graphs representing the evolution of data quality within the biodiversity domain</td>
<td>104</td>
</tr>
<tr>
<td>4.3</td>
<td>Illustration of the color-coding which represent the Record Credibility Rating for given name records</td>
<td>106</td>
</tr>
<tr>
<td>4.4</td>
<td>Paper catalog card created around the 1920s</td>
<td>109</td>
</tr>
<tr>
<td>4.5</td>
<td>Screenshot of the Aleph database back-end, which also incorporates a scan of the card catalog represented in Figure 4.4.</td>
<td>110</td>
</tr>
<tr>
<td>4.6</td>
<td>Screenshot of the Google Books interface presenting metadata of the same book as mentioned in Figures 4.4 and 4.5</td>
<td>111</td>
</tr>
<tr>
<td>4.7</td>
<td>Metadata analysis tool visualizing which Dublin Core fields are encoded on a record level</td>
<td>113</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary of the statistics calculated by a metadata analysis tool</td>
<td>113</td>
</tr>
<tr>
<td>4.9</td>
<td>Illustration of a human-readable fragment from the Dublin Core Collection Application Profile</td>
<td>115</td>
</tr>
<tr>
<td>5.1</td>
<td>Examples of problematic data at schema level</td>
<td>125</td>
</tr>
<tr>
<td>5.2</td>
<td>Examples of problematic data at instance level</td>
<td>126</td>
</tr>
<tr>
<td>5.3</td>
<td>Examples of metadata created by data-profiling to detect data errors</td>
<td>126</td>
</tr>
<tr>
<td>5.4</td>
<td>Illustration of the custom built data-profiler of Ingressus on a collection of bibliographic records</td>
<td>128</td>
</tr>
<tr>
<td>5.5</td>
<td>Report produced by the data-profiling tool of Ingressus</td>
<td>129</td>
</tr>
<tr>
<td>5.6</td>
<td>Overview of the 47 different variants of the naming of the same ethnical group in the database of the ethnographical department of the RMCA</td>
<td>131</td>
</tr>
<tr>
<td>5.7</td>
<td>Illustration of the xml file containing the specifications of which analyses are to be applied upon which fields of the metadata records</td>
<td>132</td>
</tr>
<tr>
<td>5.8</td>
<td>Graphical interface of the data-profiler</td>
<td>133</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

5.9 Illustration of the hidden whitespaces contained within the field “objectnumber” .......................................................... 137

6.1 Webpage that represents the number of catalog records that each employee has corrected within the database ............... 143
6.2 Illustration of the re-Captcha tool ........................................... 144
6.3 mBase interface which automatically detects possible erroneous values within the database ........................................... 145
6.4 Correlation between user queries and user comments .............. 152
6.5 Categories of user proposals to correct existing metadata ....... 153
6.6 Example of a user comment that adds a socio-historical context to an image ................................................................. 154
6.7 Example of a user comment relating to the user's private life .................. 155
6.8 Example of a user comment illustrating a personal opinion .......... 155
6.9 Example of a user asking a question ......................................... 156
6.10 Example of a record from the Ross Archive of African Images ...... 157
6.11 Mapping of user queries with user comment .......................... 159
6.12 Categories of user comments which complement existing metadata 160
6.13 Use of references to back up the validity of user comments ....... 161
6.14 Categories of user comments which correct existing metadata ..... 161
6.15 Example of a comment mentioning a digitally altered image ....... 163

7.1 Screenshot of the Last.fm application illustrating the services built on top of automatically inferred user needs .................... 169
7.2 Screenshot of the iGoogle interface, allowing users to “drag and drop” information of choice upon their homepage ............. 170
7.3 Homepage of the September 11th Memorial and Museum website 172
7.4 Homepage of the CollectiveAccess project website .................. 173
7.5 Illustration of how the dynamic search interface is implemented within the collection registration database .................... 176
7.6 Detailed view on how a new search form is defined ................. 177
7.7 Example of a customized search form .................................... 178

8.1 Event-aware approach toward metadata management ............... 187
8.2 Tagcloud of the social bookmarking website Delicious, which represents the most popular tags ........................................... 189
8.3 Evolution of a metadata record from 1900 until 2009 .................. 193
8.4 The dashboard of the Indianapolis Museum of Art .................. 204
8.5 Homepage of the Library and Archives of Canada, formerly known as the Public Archives of Canada and the National Library of Canada 205
8.6 Display of a search result from a database of digitized historical newspapers ............................................................... 207
## List of Tables

1.1 Comparison of the application of the stratified time concept throughout three different application domains ........................................... 12

1.2 Small case studies presented in the introduction and the first part of the dissertation ................................................................. 31

1.3 Main case studies presented in the second part of the dissertation .......................... 32

3.1 Historical overview of the implementation of ICT within the cultural heritage sector .......................................................... 71

3.2 Differences between collection and exhibition management software ................................................................. 73

3.3 Illustration of both sides of the Rijksmuseum widget ........................................... 98

5.1 Percentage of empty fields ................................................................. 134

5.2 Ten most recurrent patterns for the date of collecting field .......................... 134

5.3 Examples of different patterns to describe dimensions ....................................... 135

5.4 Most frequent titles .................................................................. 135

5.5 Use of upper- and lowercase characters .................................................. 136

5.6 Overview of the problems and suggested actions to undertake upon the metadata fields .............................................. 139

6.1 Shatford faceted classification ............................................................. 150
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAT</td>
<td>Arts and Architecture Thesaurus</td>
</tr>
<tr>
<td>AMSAB</td>
<td>Archief en Museum voor de Socialistische Arbeidersbeweging</td>
</tr>
<tr>
<td>BELSPO</td>
<td>Federal Policy for the Sciences in Belgium</td>
</tr>
<tr>
<td>CBIR</td>
<td>Content-Based Image Retrieval</td>
</tr>
<tr>
<td>CEN</td>
<td>Comité Européen de Normalisation</td>
</tr>
<tr>
<td>CHIN</td>
<td>Canadian Heritage Information Network</td>
</tr>
<tr>
<td>CIDOC-CRM</td>
<td>Conceptual Reference Model from the International Committee for Museum Documentation</td>
</tr>
<tr>
<td>DARIAH</td>
<td>Digital Research Infrastructure for the Arts and Humanities</td>
</tr>
<tr>
<td>DCMES</td>
<td>Dublin Core Metadata Element Set</td>
</tr>
<tr>
<td>ENBI</td>
<td>European Network for Biodiversity Information</td>
</tr>
<tr>
<td>EVAMP</td>
<td>European Visual Archives Market validation Project</td>
</tr>
<tr>
<td>FGCD</td>
<td>Federal Geographic Data Committee</td>
</tr>
<tr>
<td>IBBT</td>
<td>Institute of Broadband Technology</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technologies</td>
</tr>
<tr>
<td>IRPA</td>
<td>Institut Royal du Patrimoine de l’Art</td>
</tr>
<tr>
<td>ISO</td>
<td>International Standards Organization</td>
</tr>
<tr>
<td>ITIS</td>
<td>Integrated Taxonomic Information System</td>
</tr>
<tr>
<td>MARC</td>
<td>Machine Readable Cataloging</td>
</tr>
<tr>
<td>MCN</td>
<td>Museum and Computers Network</td>
</tr>
<tr>
<td>METS</td>
<td>Metadata Encoding and Transfer Syntax</td>
</tr>
<tr>
<td>MOAC</td>
<td>Museums and the Online Archive of California</td>
</tr>
<tr>
<td>MODS</td>
<td>Metadata Object Description Schema</td>
</tr>
<tr>
<td>OAI-PMH</td>
<td>Open Archives Initiative Protocol for Metadata Harvesting</td>
</tr>
<tr>
<td>OCLC</td>
<td>Online Computer Library Center</td>
</tr>
<tr>
<td>OCR</td>
<td>Optical Character Recognition</td>
</tr>
<tr>
<td>OPAC</td>
<td>Online Public Access</td>
</tr>
<tr>
<td>OWL</td>
<td>Ontology Web Language</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RMCA</td>
<td>Royal Museum for Central Africa</td>
</tr>
<tr>
<td>RSS</td>
<td>Rich Site Summary</td>
</tr>
<tr>
<td>SKOS</td>
<td>Simple Knowledge Organization System</td>
</tr>
<tr>
<td>TGN</td>
<td>Thesaurus for Geographic Names</td>
</tr>
<tr>
<td>ULAN</td>
<td>Union List of Artist Names</td>
</tr>
<tr>
<td>UNESCO</td>
<td>United Nations Educational, Scientific and Cultural Orga-</td>
</tr>
</tbody>
</table>
nization
VRA ............... Visual Resources Association
W3C ............... World Wide Web Consortium
XML ............... eXtensible Markup Language
Chapter 1

Introduction

Contents

1.1 General introduction ........................................... 2
    1.1.1 Premise defended within the dissertation .............. 2
    1.1.2 Relevance of the research .................................. 3
    1.1.3 Definition of the key concepts ............................ 4
1.2 Theoretical framework ........................................... 8
    1.2.1 Introducing the stratified time concept ................. 10
    1.2.2 Applying the stratified time to a concrete example . 11
    1.2.3 Interaction between the different time levels .......... 17
1.3 Working method .................................................. 18
    1.3.1 Literature overview ......................................... 19
    1.3.2 Case studies .................................................. 26
1.4 Structure of the dissertation ................................... 32
1.5 Original findings and outcomes ............................... 36

Summary: The introduction presents the premise which is defended within the dissertation and defines its four key concepts. The theoretical framework is then presented, which is built upon the concept of the stratified time, developed by Braudel and Boydens. The working method delivers an overview of the consulted literature and the case studies from the dissertation. The overall structure of the dissertation and the interaction between the different chapters is then explained, before a preview on the original findings and outcomes closes up the introduction.
CHAPTER 1. INTRODUCTION

1.1 General introduction

1.1.1 Premise defended within the dissertation

In contrast to most opinions, I defend that the application of ICT has not always led to better documentation practices within the cultural heritage sector. Innovative technologies offer new possibilities for metadata creation and management, but can also have a negative impact upon the quality of metadata. This observation can be framed within the larger underlying theme of how a tool influences the content it creates. Cultural heritage metadata provide one of the best possible application domains to underline this premise.

Throughout the dissertation, the core notion of metadata is understood in its widest possible sense, as data about data. The quality of metadata is defined in accordance with the ISO 9000 definition, which refers to the fitness for purpose of an object or a service [79]. This topic is studied within the specific context of the digitization of cultural heritage, which consists of the remediation of the holdings of libraries, archives and museums from an analogue to a digital medium.

From a chronological point of view, the dissertation will focus on the last decade, from the start of the first large scale digitization projects at the end of the 1990s until now. The history of ICT implementation within the cultural heritage sector, which started at the end of the 1960s in the United States, has already been briefly described. This study therefore focuses on the most recent developments of the last decade, which is characterized by the initial enthusiasm surrounding the use of databases and the Internet in the cultural heritage sector, but also the modest and short-lived outcomes of digitization projects.

Finally, the research also needs to be geographically positioned. The subject of digital cultural heritage is a truly international research topic. Scientific journals, mailing lists, forums and blogs connect an international public of practitioners and academics interested in this topic. Nevertheless, national differences do exist. USA, Australia and in lesser extent the United Kingdom have

---

1. Since the arrival of ICT in the cultural heritage domain in the 1960s and until now, the majority of professionals working in the cultural sector have hailed the new possibilities offered by automatization for the retrieval and management of cultural heritage, without reflecting on the possible negative outcomes. The proceedings of the 1968 conference "Computers and their potential applications in museums" provides valuable insights on how researchers and practitioners of the time planned to use computers for collection description. The article "The museum computer and the analysis of artistic content" by William Paisley delivers a typical example of utopian beliefs in the unlimited possibilities of using computers to search and retrieve information on cultural heritage [114]. The introduction of the Internet led to a second wave of enthusiasm, which is illustrated in the publication of "The wired museum" with for example the article by Howard Besser [14]. Other illustrative examples include "The Difficult Report" written for the Directorate-General Information Society of the European Commission [59].

2. The term “tool” will be used within this dissertation in its broadest sense, and can both equally designate a paper based catalog or an online database.

3. The term “remediation” I apply here refers to the work of Bolter and Grusin, who define the concept as “the formal logic by which new media refashion prior media forms” [19].
shown a specific interest in the empowering character of new technologies to enable user interaction with cultural heritage. The European Union has embraced ICT to enable interoperability between local European cultural heritage institutions. As most recent research focused on developments in the US, this study will give particular attention to Belgium and Europe.

### 1.1.2 Relevance of the research

From the initial enthusiasm for digitization that struck the cultural heritage sector around the end of the nineties until now, no critical analysis or debate has been held to evaluate the long term value of digitization projects. Since the end of the 1990’s the sector has embraced the Internet and digitization techniques to “mobilize” its collections. The last decade has offered a panorama of increasingly complex and ambitious projects, which were often accompanied by a general feeling of enthusiasm and belief in the advantages of Internet and databases. Just like the concepts of “democracy” and “environment” which everyone seems to think of as intrinsically good, there was, and still is, a large consensus within the professional field on the fundamental interest of the digitization of cultural heritage.

This is all the more strange when the large number of unsuccessful projects is taken into account. The goals of remediating our heritage, such as access and preservation, are often not achieved. Fact is that very few people have a clear vision on where exactly the digitization fever is leading us to. After the first decade of experience of digitization within the cultural heritage sector, time has come to reflect critically on the outcomes of the digitization fever. We need to examine the exact motives and the possible outcomes of the excitement over digitized cultural heritage. The online publication of cultural heritage is all too often considered as an end in itself, under the general vague idea of facilitating its access and use. But is this really the case?

The cultural sector does not have a tradition of self-reflective evaluation and tends not to confront initial expectations with the final outcome of projects. The term itself, “digitization project”, reveals in fact the very problematic short term quality of digitization activities. In the light of the considerable financial and organizational efforts invested into the digitization of cultural heritage, a critical analysis of the digitization discourse within the cultural heritage sector is badly needed.

Apart from reflecting on the outcomes of current and past projects, it is crucial to gain a better understanding of what the long-term consequences of the current digitization practices will be. This dissertation will demonstrate that behind the apparent straightforwardness of digitization, a vast and complex phenomenon is hiding. As new tools and technologies appear and are adopted, they change and affect the domain in which they are applied. Ingrid Mason notes:

---

4 Mobilization is used here as in the concept of “immutable and combinable mobiles” which has been defined in “Science in action” of Bruno Latour [90].
The philosophical shift from a physical to a virtual arrangement of collection, preservation, and access to cultural heritage revisits old and new social and cultural practices and assumptions, and sociopolitical power structures. What is socially acceptable or commonplace in the physical world can become enormously complex to recreate in the virtual. What has been inhibiting in the physical world can become enormously simple to enable in the virtual [101, p. 228].

1.1.3 Definition of the key concepts

The dissertation is built up around four central notions: the quality of metadata in the context of the digitization of cultural heritage. The following sections will explain in detail how these four notions are defined and understood in this work.

1.1.3.1 Defining metadata

At the most elementary level, metadata are defined as “data about data”,5 This basic definition is often detailed by referring to the structured nature of metadata and/or their machine-readable character [108].

In the contrary of these more detailed definitions, which limit metadata to structured and machine-readable data, metadata are defined within the context of this dissertation as data that have been developed throughout the history of our heritage institutions to document their collections. Following this definition, a 19th century notebook that describes an object in a narrative, unstructured way and a highly marked-up bibliographic record both correspond to my definition of metadata. Section 3.3 will elaborate on this issue by demonstrating the continuum that exists between current metadata practices and traditional indexing and cataloging.

Central within this understanding of metadata stands the idea that they allow to stabilize and combine resources that have been “mobilized” by digitization. Without the proper metadata, digitized resources lose their meaning, just like Bruno Latour mentions that “fossils, rocks or skeletons can become meaningless once in the basement of museums, if not enough context is attached to them”[90].

The notion of change, accentuated throughout the entire dissertation, is inspired by the work of Isabelle Boydens. In “Informatique, normes et temps”, she broadens the concept of metadata not only by referring to the upper-level character of metadata, but also by explicitly incorporating the notion of change in the process of documentation:

La particule grecque “meta” signifie aussi bien l'idée d'un niveau de généralité supérieur (métaphysique, métalangage,...) que celle

5See for example Jane Greenberg [67, p. 19] and Isabelle Boydens ([25, p. 107]. Zeng and Qin also give an extensive overview of the different types of metadata definitions in [163].
1.1. GENERAL INTRODUCTION

de changement (métabolisme, métastase, métépsychose,...). Cette seconde acception est également pertinente en vue de définir un système de “mét-information”, lequel implique nécessairement un processus de d’interprétation et donc de transformation de l’information documentée[25, p. 107].

The idea of change does not merely apply to the moment of the creation of the metadata, when a document is analyzed and interpreted. Once the metadata are created, they continue to evolve, and so does the described document and the world that surrounds it. In the specific context of cultural heritage institutions, where collections are gathered and documented over long periods of time, we are particularly confronted with the phenomenon of legacy metadata. Data standards, hard- and software become obsolete over time, which leads to a continuous migration process of existing metadata from one to another technological environment. Therefore, the development of a genealogy, or line of descent, of metadata takes a central role within the thesis. Based on concrete examples, the dissertation will illustrate how the construction of a family tree or “stemma codicum” of a collection registration database offers the needed framework to evaluate the quality of the metadata it holds. Section 1.2.2 will illustrate the notion of the stemma codicum by sketching the biography of the catalog from the city museum of Tienen.

1.1.3.2 Defining metadata quality

Quality is an even less tangible notion than metadata or digital cultural heritage. Boydens refers to three different interpretations of the concept of quality[25]:

- **Common understanding**: the distinctive attribute or characteristic possessed by someone or something

- **Philosophical context**: refers to the sensible aspect of someone or something, as opposed to quantity

- **Specialized meaning**: the degree of excellence of someone or something

The third interpretation is operationalized here, which immediately brings up the question of how the degree of excellence can be evaluated. This leads us to the ISO 9000 definition for quality, which states the following: “The totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs” [79]. This definition is heavily referred to in the literature, and commonly abridged as “fitness for use”. Based upon this general idea of the adequacy of use, different types of research on metadata quality have been developed from the 1990’s onwards. What is to be considered as metadata and not merely as data is a subjective and human choice, in

---

6For an introduction to the topic of the stemma codicum, see [124].

7See for example [25, 28].
function of a specific need. As a consequence, techniques and methodologies used to evaluate the quality of information or data, can also be operationalized for metadata quality research.

1.1.3.3 Defining cultural heritage

Within the framework of the dissertation, cultural heritage is defined in a very pragmatical manner as the items and collections held by libraries, archives and museums, and the documentation of which is the central focus of the dissertation. It is a deliberate decision to not extend this basic definition. The highly evolving character of what is considered as cultural heritage renders its definition extremely tedious and therefore is considered to be out of the scope of the dissertation.

Gielen and Laermans described the different evolutions the notion of cultural heritage has gone through the last two or three decades [60]. The professional community of librarians, archivists and museum conservators has been joined by non-professionals such as family tree researchers and cultural minorities in the process of defining the cultural heritage canon. Different examples and case studies will present how these small, splintered groups are actively taking part in the construction of their heritage. The dissemination of cultural heritage also increasingly differs from the traditional historiographic research. The mental frameworks that used to encompass history, such as the national history of a country or the object-centered way of organizing historical exhibitions, are more and more abandoned and make way for a more pluralistic, non-scientific approach. This tendency has been picked up by policy makers at different levels. On the website of UNESCO one can read that “the concept of heritage in our time is an open one, reflecting living culture every bit as much as that of the past”. The case study of the September 11th Memorial and Museum, presented in Section 7.3.1, is one of the world’s best examples to illustrate these evolutions of the contemporary understanding of our cultural heritage.

Apart from the enlargement of cultural heritage with elements from the popular culture, another crucial evolution of recent years is the lowering of the barriers between the three traditional cultural heritage players. Libraries, archives and museums each have their own deeply-rooted traditions regarding metadata creation, but are nonetheless more and more encouraged to collaborate more closely. A recent report from the OCLC research program gives an overview of this tendency [165]. The central motivation for bringing together these institutions is to break down the barriers between collections, which are perceived as artificial from the user point of view.

---

8As a Gartner report on metadata states: “ultimately, all metadata is data: it is the usage of the data and its context that often determine whether it is metadata or not” [16].
1.1. GENERAL INTRODUCTION

1.1.3.4 Defining the digitization of cultural heritage

Commonly, digitization is understood as the process which transfers the content of an analogue to a digital medium. Throughout the history of art, technologies have been used to represent and communicate the content of cultural heritage. With the rise of each new technology, unforeseen possibilities emerged that offered new ways to access and interpret art. However, these technologies are by no means neutral tools and do not merely meet needs. They are actively implicated in the defining and redefining of practices of heritage holders and their public [36].

The image represented in Figure 1.1 illustrates the richness but also the complexity of the different technological “layers” that can affect the content of a cultural heritage resource. The subject of the image is a painting by Rembrandt created in the mid 17th century. In the 1930’s, the Dutch photographer Bernard Eilers, who was one of the first to experiment with color photography, took a photograph of Rembrandt’s work. By using three different glassplates that each represented one of the RGB colors (red, green, blue), he managed to create a color image of Rembrandt’s painting. The digital file that was used to include the image within this dissertation was produced by scanning the three original glassplates separately. The three resulting files were then digitally combined using the layers functionality of Photoshop to re-create an “authentic” copy of the photograph Bernard Eilers took in the 1930’s.

Figure 1.1: Scan of three colored glassplate negatives made from a Rembrandt painting
CHAPTER 1. INTRODUCTION

This image illustrates in a very direct way a different set of technologies used to represent and give access to the same cultural resource. The new possibilities, but also the limits, of each technology influence the content itself. One of the most obvious illustrations is for example the relationship between the slow shutter speeds of early analogue photography and the extremely static impression that persons have on portraits. The technical requirement of staying immobile for several seconds implied that portrayed people often made an unnatural impression.

The actual digitization of content is only one part of the larger context which surrounds digitization projects. Few institutions have managed to develop a holistic approach to deal with the diverse activities required for the successful implementation of a digitization project. Figure 1.2 from the digitization tutorial of the university library of Cornell\textsuperscript{10} gives an overview of the different competencies needed. The sum of the different processes that constitute the digitization cycle require competencies of a wide range of professionals, ranging from technical to scientific skills. In practice, the uncertainty of what exactly can be considered a digitization project results in confusion and the lack of a long term vision on digitization.\textsuperscript{11}

1.2 Theoretical framework

Different strategies have been developed to approach the problem of data quality in information systems from different application domains. Three major currents can be distinguished: 1) system-centered, 2) user-centered and 3) hermeneutic approach.

The work of Thomas Redman represents the system-centered approach. Three complementary analyses build a framework for the management of database quality: data modeling, data values and data representation [123]. The user-centered approach, represented in the work of Bovee [21], focuses on the information user, by using four main criteria for data quality: accessibility, interpretability, relevance and integrity. Especially the first attribute, accessibility, receives central attention, as all other criteria are irrelevant if data are inaccessible for users. A mixed approach between the systems- and users-centered strategy was initiated in the beginning of the 1990s at the MIT, with the research of Mark Hanzen and the development of the “Total Data Quality Management” program of Richard Wang.\textsuperscript{12}

\textsuperscript{10}The tutorial is available on http://www.library.cornell.edu/preservation/tutorial/contents.html.
\textsuperscript{11}Within the Royal Museum for Central Africa for example, very diverse activities tended to be grouped under the common denominator of digitization. This situation gave rise to communication problems amongst colleagues and the different departments, but most importantly resulted in a opaque view of the aims of digitization and how/why it should be financed and with which budgets. When a consultancy report regarding the digitization policy of the museum was presented to all the head of departments and the scientific staff, the biggest discussion was on which activities should be considered as projects and which ones as part of the institutionalized activities [150].

\textsuperscript{12}See the master thesis of Mark Hansen, “Zero defect data: tackling the corporate data quality problem” [71] and the website of the Total Data Quality Management program at
1.2. THEORETICAL FRAMEWORK

Figure 1.2: Scheme representing the interdependencies between digitization goals, resources, and processes.

Quite radically distancing herself from these two research models, Isabelle Boydens has developed a hermeneutic approach toward database quality.\(^\text{13}\) Although Redman proposes an interesting model that can be applied in a wide variety of contexts, he only deals with errors that can be formally identified and measured as errors, such as incoherences and incompleteness. But nowhere Redman deals with the difficulty of the human interpretation of information.

Taking her critique a step further, Boydens especially denounces the “Data quality research” from MIT, as it is based on the assumption of the possible detection of a formal error within a database by comparing the content of the database with the reality the database strives to represent:

... Qu'est-ce qu'une donnée correcte? La réponse semble évidente: afin de vérifier la correction d'une base de données, il suffirait de comparer les valeurs qu'elle répertorie au réel observable correspondant. Une telle approche repose toutefois sur l'hypothèse implicite selon laquelle il existerait, en l'absence d'erreurs formelles, une projection biunivoque nécessaire entre une base de données et les réalités observables qu'elle représente. Or, nous soutenons que dans les domaines d'applications empiriques, objets de toute base de données, un tel isomorphisme n'existe pas[25, p. 9].

http://web.mit.edu/dqm/.

\(^{13}\)Isabelle Boydens presents an extensive review and critique upon Redman's approach. See [25, p. 52-58]. Within her approach, the notion of time takes a central role[25, p. 20-21]. This will be illustrated throughout the dissertation with the use of the stratified time concept.
CHAPTER 1. INTRODUCTION

Starting from this critique, Boydens constructs an original approach, based upon methodologies and concepts drawn from the humanities and more in particular, history. Hermeneutics, also known as the method or theory of interpretation,\(^{14}\) provide a framework for the analysis of empirical human or social realities, where deterministic approaches have no sense. The understanding of a phenomenon and its theorization evolves along with the interpretation of individual empirical facts it allows to understand. The idea of the hermeneutic circle refers to this circular way of sense making. Boydens' use of hermeneutics has been developed within the specific context of administrative databases for the management of social security, but can be generalized to other application domains, as this dissertation will illustrate.

1.2.1 Introducing the stratified time concept

In her study of administrative databases, Isabelle Boydens introduces the notion of the “temporalités étagées” to grasp the evolutions throughout time of metadata and their quality. Originally developed by the medieval historian Fernand Braudel, this concept of the stratified time provides a framework to identify a hierarchy between different types of co-existing evolutions.\(^{15}\)

Three different time zones are postulated: the “temps long”, the “temps intermédiaire” and the “temps court”. Braudel associated the “temps long” within his monumental history of the Mediterranean area during the reign of Philip II with geographical dispositions such as the climate. Isabelle Boydens places here the evolution of the legal norms, which has a clear impact on the structure of the administrative databases she studied.\(^{16}\) While Braudel associated the economical evolutions with the “temps intermédiaire”, Boydens places here the transformation of the technological and administrative apparatus. And lastly, while the daily reality of the political actions belongs for Braudel to the “temps court”, Boydens situates here the evolution of the “real”, which is the object of the legal norms and the technological/administrative apparatus.

I therefore explicitly do not approach metadata quality solely by relying on a conceptual framework or by trying to operationalize an innovative technology, but instead propose an interaction between both. With regard to the framework, I adapted Isabelle Boydens' hermeneutic approach to my specific context of metadata creation within the cultural heritage sector, in order to distinguish the different levels of evolution that each have an impact upon metadata quality:

- **Temps long**: evolution of the policies toward cultural heritage preservation and access, including the shifts in the perception of what should be considered as cultural heritage and how the public should relate to


\(^{15}\)Braudel elaborated the concept of the stratified time within his seminal work “La Méditerranée et le monde méditerranéen l'époque de Philippe II” [27].

\(^{16}\)Boydens presents her adoption of the stratified time concept in detail in [25, p. 170-175].
that heritage. These evolutions were analyzed in the research of cultural historians such as David Lowenthal [93] or cultural sociologists such as Pascal Gielen [60]. I also followed the effect of mentality changes by analyzing the policies of funding agencies such as the Directorate General Information Society of the European Commission, the Federal Policy for the Sciences in Belgium (BELSPO) and the Institute of Broadband Technology (IBBT) in Flanders. These evolutions will be discussed in Chapter 2.

- **Temps intermédiaire**: evolution of the technologies, in the broad sense of the word, used for metadata creation. These have evolved from unstructured notice books in the 19th century to the highly structured relational databases currently in use. Descriptive tools such as classification schemes and thesauri also fall under this category. Apart from an analysis of the historical evolution of these technologies that will be illustrated with concrete examples from the terrain, the evolution of classification and indexing schemes is also discussed. Geoffrey Bowker is one of the only scholars that worked on these issues [24, 23]. This level correlates with the content of Chapter 3, where a historical overview of the documentation practices in the cultural heritage sector is given.

- **Temps court**: evolution of the metadata themselves, which are the object of the policies and the technological apparatus. The concrete analysis of metadata, with the help of data-profiling techniques, is at the centre of Chapter 5. Chapter 6 will illustrate how collaborative practices have a concrete impact upon metadata quality. Chapter 7 will then analyze how the day-to-day use of metadata can lead to changes within the schema of metadata, discussed within the intermediate time.

Table 1.1 presents an overview of the three different uses of the stratified time, according to each specific application domain.

1.2.2 Applying the stratified time to a concrete example

There is no better way to illustrate the interest of the stratified time concept than to take a real-life example of a catalog that has known important transformations throughout its existence. This section will therefore sketch the historical evolution of the catalog of the museum of the city of Tienen.\(^{17}\)

1.2.2.1 Historical overview of the cataloging processes

Figures 1.3 and 1.4 give an overview of the different cataloging tools used throughout the history of the museum, which was created in 1898. Two years later, the council of the museum published the “Inventaire des objets appartenant au musée communal”. This catalog consists of a chronologically ordered

\(^{17}\)The history of the catalog of the museum is based upon unpublished internal documents from the museum and an interview with Staf Thomas, the current director of the museum.
CHAPTER 1. INTRODUCTION

<table>
<thead>
<tr>
<th>Application domain</th>
<th>Current approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>Administrative databases</td>
</tr>
<tr>
<td>Long time</td>
<td>Evolution of the climate and the geography</td>
</tr>
<tr>
<td>Intermediate time</td>
<td>Evolution of the economic situation</td>
</tr>
<tr>
<td>Short time</td>
<td>Evolution of the political situation</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of the application of the stratified time concept throughout three different application domains

list, which includes for each object a concise title describing the content of the object, which occasionally includes a date, as for example “Diplôme reconnais- sant les armoiries de la ville, 1841”. Today, there are still some objects in the depot which have an inventory number referring to the number the object received within this catalog.

Since 1910, the city archives no longer mention the existence of the city museum. The collection is brought again to the attention by Jan Wauters in 1939 who launched the “Folkloristisch en Oudheidkundig Museum”. Wauters created a handwritten and very concise inventory of the collection in 1942. On May 25th, 1944 the allied troops bombed the building where the collection was housed, and the objects were relocated in a variety of locations. This motivated Wauters to quickly draw up another handwritten inventory in a notebook, but which now contains a more detailed description of the collection. No evidence of the numbering that was used in this notebook can be found in the depot today.

In the 1960s, P. Dewalhens created an exhibition based upon the collection, and in 1963 published a new inventory entitled “Inventaire des tableaux de peintres belges des XIXe et XXe siecles exposés à l’Hôtel de Ville de Tirlemont”. For the first time, this inventory included formal characteristics of the objects, such as its dimensions.

Then finally in 1977 Staf Thomas was appointed as archivist/conservator, and was commissioned to create a city museum. His first task was to re-establish an inventory. The museum used pre-printed paper inventar cards based upon the model used by two large museums in Brussels. A field that was considered to be important was the series of abbreviations printed on the top of each card. This field was inspired by an approach used within the insurance world, but was never actually implemented as the technology did not break through on a sufficiently large scale.
1.2. THEORETICAL FRAMEWORK

The next big step was then taken in 1988 when the museum decided to use FileMaker Pro to document its holdings. The metadata schema created within the database was based upon the registration model presented in the publication of Jeanne Hogenboom [77]. This is the model still used today.

In 2008, the province of Vlaams-Brabant launched the idea of creating a common database for cultural heritage, in order to enable the public to consult all of the digital cultural heritage collections of local and regional museums and archives through a single website. The city museum of Tienen would like to participate in this project but before an actual collaboration can start by migrating their metadata to the common database, major efforts need to be put into the quality of the metadata.

It is clear that FileMaker Pro is not an ideal software solution for collection description, and therefore the museum is currently investigating the possibility of moving toward another software solution. The open source collection management system CollectiveAccess seems to fit the needs of the museum. Before the existing metadata can be moved to a new software platform, an analysis needs to be done of the quality of the metadata, which will probably lead to an adaptation of the current metadata scheme and a clean-up operation of the poor quality metadata.

1.2.2.2 Applying the stratified time on the preliminary case study

The different evolutions described above will be mapped onto the stratified time canvas:

- **Temps long**: Two different long term evolutions can be identified. The first catalog started out as a personal initiative of the head of the collection, with the purpose of publishing an inventory of the objects from the museum collection to illustrate the cultural heritage of the city. In the following decades, the growing professionalization of documentation practices is reflected within the use of pre-printed catalog cards. The implementation of FileMaker Pro illustrates the need for an internal instrument to monitor and manage the growing collection. This focus on the internal use of documentation has gradually shifted since the end of the 1990s to an emphasis on the relevance toward the public and a collaboration between different heritage institutions. The administration of the Flemish ministry of culture has been pro-actively supporting collection registration by financing digitization projects and is now promoting the interoperability in between individual collections.

Apart from this influence of the government funding policies upon metadata quality, the gradual evolution of the use from French to Dutch as an

---

18 For the background of the project see http://nieuwsbrieven.vlaamsbrabant.be/vrije-tijd-cultuur/cultuur/erfgoed-en-monumentenzorg/erfgoedbank/index.jsp
19 See http://www.collectiveaccess.org, Chapter 7 will give more information regarding this software.
20 A concrete example of how the public sector promotes metadata creation can be found on http://musea.oost-vlaanderen.be/public/index.cfm.
CHAPTER 1. INTRODUCTION

Figure 1.3: Overview of the different cataloging tools used throughout the history of the city museum of Tienen, part 1
1.2. THEORETICAL FRAMEWORK

indexing language is also an important long term development to analyze. This evolution stands completely independent from the cultural sector, but
has a tremendous impact on how the metadata can be consulted.\textsuperscript{21}

- **Temps intermédiaire**: Evolution from a paper-based catalog, to the stand-alone FileMaker Pro with the metadata scheme inspired by the cataloging standard of Hogenboom, and then to the web-based CollectiveAccess which will use the metadata scheme of the Erfgoedbank Limburg.

- **Temps court**: The existing 1,000 descriptions were expanded gradually from 1977 onwards to 6,680 records within the database. Until recently, the head of the collection created all of the descriptions. This year an extra employee was hired for a two year period to work specifically on the catalog.

The evolution of the metadata can be illustrated when using the example of the description of one specific object. Below I have summed up the different variations of the metadata as they were represented by the different technologies as described above:

- Printed catalog from 1900: “20. Un mortier moyen modele en bronze portant l’inscription Petrus Vanden Gheyn me fecit 1546. Le pilon manque”
- FileMaker Pro database, with a metadata scheme based upon printed catalog sheets from 1977, which contained 27 metadata fields that could be filled out. The following fields are used for the description of our object:
  1. Plaatsnaam: Tienen - Toreke - 328
  2. Postnummer: 3300
  3. Inventarissenummer: 1977/0051
  4. Object trefwoord: apotheek, vijzel
  6. Datering: 1546
  7. Materiaal trefwoord: brons
  8. Toestand trefwoord: 9
  9. Maten: h. 11,3 cm, diam. bovenaan 15,2 cm
  10. Bewaarplaats: Bureau
  11. Kaart ingevuld door: Jos

\textsuperscript{21}Section 5.4 will present the case study of the Royal Museum of Central Africa, where the use of both French and Dutch results in information retrieval problems.
1.2. **THEORETICAL FRAMEWORK**

12. op: 08 april 1997

- This record was corrected a year later by the director of the museum. The following fields were corrected:
  1. Object trefwoord: mortier, vijzel
  3. Toestand trefwoord: 7
  4. Maten: h. 11 cm; diam. voet 9,3 cm
  5. Bewaarplaats:
  6. Kaart ingevuld door: Staf
  7. op: 11 mei 1998

### 1.2.3 Interaction between the different time levels

The three different time levels as originally laid out by Fernand Braudel interact in a cascading way. The long time, with the climatological and the geographical conditions of a region as central elements, influences the two lower levels. The economical situation, represented in the intermediate time, has then in its turn also an impact on the political life, of which evolutions are analyzed within the short time. Within the original framework of Braudel, the interaction is clearly a one-way process.

A more complex situation occurs when transposing Braudel’s framework to the context of information systems. As within the original context of Braudel, the longer term levels influence the shorter term levels. Evolutions in the legislation on social security for example have to be reflected in the structure of the databases which administrate the social security. Changes within these databases then have a real-life impact on how the social security of citizens is managed.

However, Isabelle Boydens noticed that the interaction between the three different levels is not limited to a one-way interaction. New technological developments, which belong to the “temps intermédiaire”, can affect the “temps long”. Building further on the example I just mentioned, the development of a system for digital autographs led to changes within the legislation, which has to approve this new way of authenticating documents. Changes of the object of documentation itself, analyzed within the “temps court”, alter the two superior levels. The growing number of single-parent households and recomposed families for example change the characteristics of the citizen, the object documented within administrative databases. This new reality is slowly embedded within the legislation which needs to reflect the changing nature of society. In its turn, the new
realities which are acknowledged by the legislator are then incorporated
within the structure of the social security databases. To acknowledge this
interaction between the different hierarchical levels, which differs from
the original concept of Braudel, Isabelle Boydens incorporated the notion
of the “continuums évolutifs”.22

The same interdependence between the different time levels also occurs
with the use of the stratified time within my application domain. Men-
talties and policies from the long time influence both the development of
standards and technologies, and the concrete creation of metadata in our
cultural institutions. The impact of the long upon the short time can take
place in an indirect manner through the intermediate time. The long time
can also have an immediate impact upon the short time. The growing
importance of the project-based way of financing digitization projects by
authorities at different levels (regional, national or European) can have
a negative effect on the quality of metadata. The short term character of
the employment of metadata practitioners hired to work on digitization
projects often results in a lack of involvement and understanding of the
collections they are describing. Section 3.4.2 presents an analysis of this
situation and its impact on the professional status of metadata practition-
ners.

By analogy with how recomposed families affect legislation and the struc-
ture of administrative databases, the lower levels can also have an impact
upon the superior ones within the context of cultural heritage databases.
Chapter 7 for example will illustrate how the interaction between users
and the existing metadata within the search and display interface, that
takes place in the short time, can influence the re-design of the structure
of the metadata.

1.3 Working method

Two main characteristics of my research domain have led to a specific
working method to defend and elaborate the premise of this dissertation.
First of all, the absence of an established and clear defined epistemological
framework on digital cultural heritage obliged me to consult the meth-
ods and literature of other domains such as information and computer
science. Furthermore, as I already demonstrated in Section 1.2.2 of the
dissertation, the topic of metadata quality within the cultural sector is by
its nature very empirical. The dissertation therefore confronts the theo-
retical findings with a number of case studies. The following two sections
present both the sources from which the theoretical findings are drawn
and the different case studies that intervene throughout the research.

22The concept of the “continuums évolutifs” has been originally elaborated by Norbert Elias. See
1.3. WORKING METHOD

1.3.1 Literature overview

1.3.1.1 Specific works on the topic of the dissertation

As the research on metadata quality of digital cultural heritage is a recent research domain that emerged only three to four years ago, there are no well-established research communities that each have their own distinctive approach. I therefore can not position myself against currents that regroup several researchers.

A small number of papers and book chapters have been published on the specific topic of metadata quality in the cultural heritage industry, all of which were published in either 2007 or 2008. In 2008, two books appeared on metadata in the context of digital libraries, which also attribute attention to the issue of metadata quality. Marcia Lei Zeng and Jian Qin provide an extensive overview of the existing research on metadata quality [163]. Muriel Foulonneau and Jen Riley included in their book a section on the quality control for metadata creation, where they rightly point out the iterative character of metadata quality evaluation [56]. Both books especially focus on the issue of metadata quality within digital library repositories that contain metadata from multiple sources.

The paper of Ochoa and Duval [109] proposes an automatic evaluation of metadata quality. Their approach is valuable due to the fully automated manner in which the evaluation is performed. However, their method is based upon a comparison between the actual content of the metadata record. They state: “The correctness and, thus the accuracy, could be considered as the semantic distance between the information that a user could extract from the metadata record and the information that the same user could obtain from the resource itself. The shorter the distance, the higher the accuracy of the metadata record” [109]. I would like to mention two fundamental objections. Their method to detect how many words match between the metadata record and the resource it describes can first of all only function for machine-readable textual documents. More importantly, however, the underlying idea that the content of a metadata record literally reproduces the content of a resource is a very reductionist vision on the methods and goals of indexing and cataloging. See Section 1.3.1.3 for a more detailed review of the paper.

Another publication which addresses the topic of this dissertation is the bookchapter Ingrid Mason wrote for the first academic monograph on the digitization of cultural heritage [101]. Although the chapter delivers an original insight into how metadata standards are constructed within the cultural heritage community, the limitations and problems of the actual metadata practices on the terrain are not sufficiently acknowledged.

The first academic journal on digital cultural heritage, the ACM Journal on Computing and Cultural Heritage (JOCCH) was just launched in June
2008. Metadata is one of its core topics, but the first issue did not contain articles specifically on metadata quality.\(^{23}\)

In general, most of the consulted literature either deals with metadata or with the digitization of cultural heritage. These two subtopics are studied within different academic disciplines, which each have their specific focus from the point of view of the dissertation. The next few sections give an overview of the different disciplines. A summary at the end of each section provides a short overview of which aspects of the academic discipline were used to build the framework of my dissertation and which were the shortcomings to answer the specific research questions.

1.3.1.2 Information science

Information science is the obvious scientific discipline from which to study metadata quality. However, the definition and the naming of the discipline itself is highly problematic. Several authors, such as \(^{[164, 10, 125, 122]}\) have published extensively on the identity question of the discipline.\(^{24}\) The boundaries between this discipline and computer science also form the object of a vivid debate, but this matter is out of scope of the dissertation.

The research tradition within information science on cataloging and indexing rules, the creation of thesauri and taxonomies for example provides the ideal framework for studying metadata quality. Furthermore, this discipline has always worked in close collaboration with heritage institutions to study descriptive practices and to educate future employees of libraries, archives and museums. From an international point of view, the School of Library and Information Science at the University of North Carolina at Chapel Hill\(^{25}\) and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign\(^{26}\) have some outstanding scholars who focus on metadata research.

The following set of publications from information science inspired the dissertation in an important manner. First of all, the monograph of Isabelle Boydens “Informaticque, normes et temps” \(^{[25]}\) provides the conceptual foundations upon which my specific notion of metadata quality was constructed.\(^{27}\) The chapter Diane Hillman and Thomas Bruce wrote on metadata quality in the book “Metadata in practice” presents the first framework for assessing metadata quality \(^{[28]}\). The proceedings of the Inter-
1.3. WORKING METHOD

national Conference on Dublin Core and Metadata Applications $^{28}$, with contributions such as papers on the link between metadata quality and application profiles [76], regroup a number of articles on my topic.

However, as Tefco Saracevic points out, the discussion on quality assessment within information science preceeds and outnumbers the actual evaluation work itself, as it was the case with the assessment of information retrieval in the 1960s [126]. Saracevic notes that the research falls in two main categories: 1) works which discuss and propose evaluation frameworks and guidelines, and 2) studies presenting case studies where analyses are performed upon actual data. The second category can even be split up in two subcategories: work based upon 1) hard data, which imply quantitative methods on substantial datasets or 2) soft data, based upon small samples, which are treated manually.

The large majority of research focuses on evaluation frameworks and guidelines. Within the minority of publications that look at actual case studies, most manually analyze small samples. Only [109] and [54] present case studies with large test collections of metadata which are assessed in an automated manner.

The position of my research against the existing metadata quality research from information science can be summarized in the following points:

- **Interests:**
  - Provides a solid conceptual framework for the analysis of metadata quality

- **Criticism:**
  - Strong focus on library metadata, studies on museum or archival metadata are rare
  - Studies of metadata quality either remain too conceptual or they only present a “one shot” approach toward practical metadata evaluation, but never offer a generalizable method that can be deployed throughout different contexts

1.3.1.3 Computer science

The computer science domain has offered tremendous help for the creation and management of metadata within digitization projects of cultural heritage. Section 3.2 provides a short overview of the different technological evolutions that occurred and influenced the cultural heritage sector. Two different research communities in particular have provided significant progress in my application domain: Content-Based Image Retrieval (CBIR) and the semantic web community.

$^{28}$See http://www.dcmipubs.org/ojs/index.php/pubs/
Other computer science subdomains that focus on databases and data mining have also contributed to the research on metadata quality, such as work on the automated detection of errors within large metadata sets \[32\]. The rules for sound database design, which were already laid out in the 1970s and 1980s, can still be applied to tackle a number of metadata quality issues that appear in current cultural heritage databases.\footnote{The reference work of Elmasri and Navathe provides a good overview of database design \[49\].} Section 5 on data profiling will elaborate on this.

**Content-based image retrieval**

Work on CBIR has its origin in the digital imaging research, stimulated in the 1960s and 1970s by military and spatial developments. It is only in the middle of the 1990s that CBIR software came on to the market for medical and civil applications.\footnote{Jorgensen provides an overview of CBIR methods and projects for the automated indexing of images \[81\].} Ever since, a large number of prototypes and research projects, such as \[20, 147\], have been launched to illustrate the opportunities to automatically extract metadata from the content of the images by analyzing the shapes, patterns and colors that appear in the image. A concrete example of the integration of a content-based image search engine can be found on the website of the Russian Hermitage museum\footnote{See the museum’s website on http://www.hermitagemuseum.org.}, where IBMs Query by Image Content (QBIC) search engine can be tested out. Figure 1.5 illustrates the search interface, where users can define patterns and colors of the images they would like to retrieve.

However, the application of CBIR within the cultural heritage domain has not been very successful. The users of cultural heritage collections attach much importance to the retrieval of images by “inferred semantic content” \[50\]. The difference between the low-level semantics that can be automatically extracted by CBIR and the high-level semantics produced by human indexing is commonly referred to as the “semantic gap”. Section 2.4.1 goes further into the details of the consequences of this semantic gap for the metadata quality within cultural heritage databases.

**Semantic web**

Partly building upon the tradition of the artificial intelligence research from the 1980s and 1990s, the semantic web community was gradually built around the World Wide Web Consortium (W3C) shortly after 2000 \[13\]. This community has played a major role in the uptake of metadata as a research topic and is currently helping to bridge the gap between information and computer science. The publication that followed for example the seminar organized in October 2008 by the Institut National de Recherche en Informatique et en Automatique (INRIA)
1.3. WORKING METHOD

Figure 1.5: Illustration of the Query by Image Content search interface, as implemented in the search interface of the Hermitage state museum

entitled “Métadonnées: mutations et perspectives” includes two chapters [153, 58] which clearly link the research on metadata with the work on the semantic web. The book “Knitting the semantic web” [104] also elaborates on this link between metadata research and the semantic web, and demonstrates how the semantic web research community can learn from traditional tools and methodologies of information science and library management.

What is for now the practical added-value of the semantic web for the cultural heritage sector? The Dutch Multimedian project32 is a good example, as it regroups some of the leading semantic web researchers that focus on deploying semantic web standards such as the Resource Description Framework (RDF), the Ontology Web Language (OWL) and the Simple Knowledge Organization System (SKOS) within the cultural heritage domain. The Multimedian search interface does propose extensive search possibilities and collections items, artists and periods which are interlinked in a meaningful and rich way.

However, it is crucial to understand that these extended search and browse possibilities are not directly the outcomes of semantic web standards and

---

32See articles such as [144, 159, 128] and the project website on http://e-culture.multimedian.nl/.
CHAPTER 1. INTRODUCTION

technologies. Decades of traditional cataloging, indexing and thesaurus development provided the high-quality metadata that were needed to demonstrate the utility of semantic web standards. As is mentioned in Section 3.5.1, authors such as Cameron and Robinson [30] criticize the traditional metadata practices in favor of semantic web standards, which makes no sense as the latter rely for their success on the existing metadata collections as contained by libraries, archives and museums.

Unfortunately the CBIR and the semantic web research communities do not collaborate. The low level semantics that can be discovered through CBIR techniques could in some cases be leveraged with the help of ontologies by pointing out semantical links between low level semantics and thus discovering higher level concepts.

Below I summarize both the converging and diverging points of the discussed computer research communities with my research:

- Interest:
  - The work on metadata standardization by the semantic web community has been extremely valuable in increasing awareness on the importance of metadata standards.

- Criticism:
  - In the quest for interoperability, the focus on the local needs for metadata creation is often forgotten and interoperability is too much considered an end in itself.
  - CBIR software can currently only deduce low-level semantics. The connotative semantics that are of great importance within the cultural heritage sector still have to be manually created.

1.3.1.4 Digital humanities and new media studies

The maturation of the Internet and digitization technologies in the second half of the 1990s led to a new academic research field, which aims to grasp the nature of the new possibilities offered by ICT from a humanities point of view. A set of new but rather unstable and ill-defined research fields emerged on the crossroads of computer science with cultural studies, history, sociology and other disciplines from the humanities. New research departments and institutions were created to professionalize these new academic fields, such as the Media Lab at the Massachusetts Institute of Technology (MIT)\textsuperscript{33} or the Jan van Eyck Academie.\textsuperscript{34}

New media studies was one of the most recurring terms to designate these emerging fields at the end of the 1990s. Two publications, both published

\textsuperscript{33}See http://www.media.mit.edu/

\textsuperscript{34}See http://www.janvaneyck.nl/
around 2000 at MIT Press, were highly influential: “Remediation. Understanding New Media” of Jay David Bolter and Richard Grusin [19] and “The language of New Media” of Lev Manovich [97]. Especially the latter delivered an original view on databases, by analyzing them as the dominating cultural form of expression. The monograph of Elizabeth Eisenstein, published in 1979 on the impact of the printing press, proved to be very contemporary by the way it allowed the comparison between certain effects of the printing press and the Internet or digitization [47].

More specifically on the topic of digital cultural heritage, the 1997 publication “The wired museum: emerging technologies and changing paradigms” [4] offered a first collection of chapters on how museums could gain from digital reproduction technologies and the Internet. This publication thrived on the digitization and interactivity hype that affected the second half of the 1990s and is largely outdated now.

The follow-up book “Theorizing digital cultural heritage. A critical discourse” [30] that was published in 2007 has an ambitious title, but unfortunately does not deliver the promised solid theoretical background by its lack of coherence between the different chapters and the strong focus on highly specific and rather small, experimental case studies which are not representative for the whole sector.

The symposium “Force of metadata”, organized in November 2008 by the Goldsmiths Media Research Centre, illustrates the current interest of new media in metadata. The abstract that describes the intent of the symposium also points to the very vague manner in which this domain approaches metadata:

Metadata rules the web. Its power goes beyond merely ordering descriptions of data. Metadata administers access, predetermines preferences, enables surveillance, automates transfextuality, and shapes our experience. As metadata management becomes more and more effective and ubiquitous, it is time to ask: Are we witnessing the birth of a new regime of attention, of media control and media power?36

Again, a summary explains the position of my research against the new media studies approach:

- Interest:
  - By avoiding a solely technical focus on the subject, and by allowing a more social and historical scope, an original approach toward digital cultural heritage can be developed by comparing the effects of the Internet and digitization tools with other evolutions in the history of media.

---

35See also [96] and Lev Manovich’s website http://www.manovich.net/.
36See http://www.goldsmiths.ac.uk/media-research-centre/project2-news.php
CHAPTER 1. INTRODUCTION

– Criticism:

* Only a minority of new media publications offer a solid theoretical basis and provide an empirical framework. Many publications focus on notoriously vague concepts such as “archive” or “simulacra” from deconstructive and postmodern philosophers such as Jacques Derrida and Jean Baudrillard, but lose their meaning as a result of the very abstract level at which they operationalize these concepts.

1.3.2 Case studies

1.3.2.1 On the necessity of an empirical framework

Research which is so intrinsically attached to the empirical realm, as it is the case with metadata quality, necessarily needs to verify its claims and research results by analyzing practices from the terrain. The premise of this dissertation states that technological evolutions do not always have a positive impact upon the quality of metadata in the cultural heritage sector. Therefore, I need to verify this statement by confronting it to the reality. Case studies are in this context the most adequate methodological tool:

Predictive theories and universals cannot be found in the study of human affairs. Concrete, context-dependent knowledge is, therefore, more valuable than the vain search for predictive theories and universals. (...) One can often generalize on the basis of a single case, and the case study may be central to scientific development via generalization as a supplement or alternative to other methods. But formal generalization is overvalued as a source of scientific development, whereas the force of example is underestimated[53, p.224,228].

However, not all research on metadata quality shares this concern. The publication of Batini and Scannapieca [11] outruns itself in references to other works, schemes, typologies and conceptualizations, and loses its relevance by the absence of sufficient references to real life examples. Within my specific application domain the researchers that work on a more conceptual level tend to disregard the importance of coupling their findings with practical implementations. The work on the “event-aware” metadata model from Carl Lagoze [87] is for example fascinating from a conceptual point of view. It is unlikely that the model itself will ever be implemented within the descriptive workflow of an institution, as the model adds a complexity to metadata creation and maintenance which is difficult to support in practice.
1.3. **WORKING METHOD**

The scientific interest of including a case study within research is the ability to extrapolate the conclusions drawn from unique and specific observations to other contexts from the same application domain. In the context of the digitization of cultural heritage, a significant number of case studies refer to highly experimental prototype software applications that will probably never be applied on a large scale. Here I can specifically refer to the CBIR and semantic web research projects [111, 128]. The majority of the case studies in “Theorizing digital cultural heritage. A critical discourse” are also very experimental in nature and do not reflect the current practices and needs of the majority of cultural heritage institutions.37

I clearly acknowledge the importance of investing in innovative technologies and to evaluate and report on them by publishing case studies. I also think that there is a need for case studies that point out widely occurring day-to-day problems. Proposed solutions are then likely to be of use to a large array of institutions within the same domain.

Alternative methodologies for collecting empirical data exist, such as surveys, interviews and focus groups. The major drawback of all these methodologies is that they are very resource intensive. The report on the “Quality Metrics” project, conducted by researchers from Emory University and Virginia Tech gives an impression on the organisational difficulty of focus groups and real-time interviews with users [86]. Online surveys conducted by e-mail are less resource intensive, but research shows that they are not always effective [118]. The end report of the user needs study of the Museum and Online Archive of California (MOAC) project also clearly illustrates the huge difficulties in obtaining sufficient and trustworthy data when using online surveys, even when combined with interviews and focus groups [61]. Section 2.2 presents in more detail the difficulties that were encountered in this project.

Therefore, a more pragmatic approach is adopted; one that takes into account the economic and organizational realities of the cultural heritage sector. The operational outcome of the dissertation is a set of tools and methodologies to monitor and enhance the quality of metadata within the cultural heritage domain, which require a minimum of resources to implement. The case studies from the second part of the thesis acknowledge the above mentioned realities, by using for example only open source software.

### 1.3.2.2 Difficulties associated with the use of case studies

As the list of interviewed metadata practitioners of Section A.1 in the appendices illustrates, an important number of cultural heritage practition-
ers were interviewed when preparing this dissertation. Once the research questions were defined, the search for the best possible case studies to illustrate each of the three operational strategies was very time intensive. Between the first moment of contact with the metadata practitioners of an institution and the moment I was able to obtain a concrete export of metadata from the database manager of the same institution, a large number of interviews and meetings occurred.

Each library, archive or museum has to deal with problems in relation to the quality of their metadata. Nevertheless, collection holders are not eager to communicate about these issues and have someone external assessing the quality of their metadata. This situation obviously complicated the selection of the case studies. Apart from the understandable reluctance of having an external party evaluating the quality of documentation practices, institutions such as the September 11th Memorial and Museum can obviously not allow access to the highly sensitive and personal information from their collection registration database. I am therefore very grateful that their database manager provided me with an export of the logfiles created by the dynamic search interface, which only contained information regarding metadata fields.

1.3.2.3 Typology of case studies

Different strategies for the selection of case studies exist. Figure 1.6 gives an overview of the different options, defined by Bent Flyvbjerg, that can be adopted to select cases [53]. As Flyvbjerg notes, the different strategies are not mutually exclusive. Cases can be simultaneously critical, paradigmatic and/or examples of extreme variation case studies, as the combination of these strategies allows different perspectives [53].

Within my dissertation, different types of case studies intervene. The majority are critical case studies, as the specific cases which are discussed are of strategic importance in relation to the problems and research questions posed in the dissertation. For example, the database of the Royal Museum of Central Africa is a critical case study for the application of data profiling techniques, as the metadata within the collection registration database were developed over a process of several decades by different people. This resulted in a large number of inconsistencies within the metadata, which makes this specific case interesting to illustrate the possibilities of the data profiling tool.

Chapter 6, which focuses on the possibilities of user comments for the correction of metadata, is based upon two maximum variation case studies. These types of case studies are very different on one specific dimension. The case study of the user comments of the image database of the National Archives (see Section 6.3.2) analyzes comments that come from a very wide audience, mostly consisting of laypersons such as amateur his-
1.3. WORKING METHOD

<table>
<thead>
<tr>
<th>Type of Selection</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Random selection</td>
<td>To avoid systematic biases in the sample. The sample’s size is decisive for generalization.</td>
</tr>
<tr>
<td>1. Random sample</td>
<td>To achieve a representative sample that allows for generalization for the entire population.</td>
</tr>
<tr>
<td>2. Stratified sample</td>
<td>To generalize for specially selected subgroups within the population.</td>
</tr>
<tr>
<td>B. Information-oriented selection</td>
<td>To maximize the utility of information from small samples and single cases. Cases are selected on the basis of expectations about their information content.</td>
</tr>
<tr>
<td>1. Extreme/deviant cases</td>
<td>To obtain information on unusual cases, which can be especially problematic or especially good in a more closely defined sense.</td>
</tr>
<tr>
<td>2. Maximum variation cases</td>
<td>To obtain information about the significance of various circumstances for case process and outcome (e.g., three to four cases that are very different on one dimension: size, form of organization, location, budget).</td>
</tr>
<tr>
<td>3. Critical cases</td>
<td>To achieve information that permits logical deductions of the type, “If this is (not) valid for this case, then it applies to all (no) cases.”</td>
</tr>
<tr>
<td>4. Paradigmatic cases</td>
<td>To develop a metaphor or establish a school for the domain that the case concerns.</td>
</tr>
</tbody>
</table>

Figure 1.6: Overview of the different strategies for the selection of samples and cases, as defined by Bent Flyvbjerg

A second case study complements the first by analyzing the user comments that are sent by a restricted group of worldwide experts on African art to the Ross Archive of African Images (see Section 6.3.3).

The small case study on Magnum Photos (see Section 2.3.2.1) is the only example of an extreme or deviant case study. The self-sustainability of digitization projects within the cultural heritage sector by achieving a return-on-investment with the commercialization of digitized materials is for most institutions an unachievable goal. The case of Magnum Photos illustrates that it is possible, but more importantly, points out the impact it has on metadata quality.

1.3.2.4 Summary of the case studies applied throughout the dissertation

The emerging research field on metadata quality and more generally on the digitization of cultural heritage is a truly international research domain, in the sense that the issues have come to the forefront on all different levels of heritage institutions worldwide. Cultural institutions of all sizes and manners are starting up digitization projects and are consequently confronted with the matter of metadata quality. In order to find a maximum number of references to actual problems from the terrain, I
could not focus on one single case study.

On the contrary, I strived to incorporate examples and experiences from all possible levels from the cultural heritage domain: international, European, national, regional and local. Small local institutions are confronted with a different set of problems than the national libraries that embark on collaborative digitization projects [127, 128].

The second part of the dissertation is completely built by looking at full-blown case studies, in order to validate the three proposed metadata quality strategies. However, the introduction and the first part also contains a limited number of small case studies. Tables 1.2 and 1.3 give an overview of the different cases and the reason why they have been included in the dissertation.

Case studies from the introduction and the first part of the dissertation

The introduction and the first part of the dissertation are mainly based upon literature from different disciplines. However, I decided to include a restricted number of small case studies to clarify and detail certain elements. The historical evolution of a specific metadata record of the city museum of Tienen, which figures in Section 1.2.2, immediately renders the topic of the dissertation more comprehensible by giving a concrete example of the evolution of metadata quality throughout a set of different technologies. Both the examples of Magnum Photos (see Section 2.3.2.1) and the Google books project of the university library of Ghent (see Section 4.3.3) add a necessary complement to the conclusions drawn from the literature.

These cases are considered as small, since I did not perform a thorough analysis of large sets of data, which is the case with the full-blown case studies presented in the second part of the dissertation. Table 1.2 presents an overview of the small case studies. However in all three cases I interviewed several members of the organizations who are responsible for the management of the metadata records. This allowed me to have access to first-hand examples illustrating metadata quality issues.

Case studies from the second part of the dissertation

Table 1.3 gives an overview of the different case studies of the second part of the dissertation. Chapter 5 presents the first major case study and consists of the application of five data profiling scripts on a .csv export file of the ethnographic department of the Royal Museum for Central Africa. The export was made by the database manager of the ethnographic department, consisting of 69,719 records. Each metadata record has 13 fields (object id, object number, object count, date of collection, date of
1.3. WORKING METHOD

<table>
<thead>
<tr>
<th>Introduction</th>
<th>Name</th>
<th>Type</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>City museum of Tienen</td>
<td>Critical</td>
<td>The whole range of technologies used for documentation purposes has been used</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>throughout the history of this institution</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Magnum Photos</td>
<td>Extreme</td>
<td>In contrast with cultural institutions, Magnum can be self sufficient</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Google books project at the UGent library</td>
<td>Critical</td>
<td>The library is one of the European forerunners with the large scale digitization of their holdings</td>
</tr>
</tbody>
</table>

Table 1.2: Small case studies presented in the introduction and the first part of the dissertation

entry, date of production, title, medium, dimensions, thesaurus terms, description, old region, actual region). The metadata are in French, with some minor exceptions in Dutch. The data profiling tool that performs the scripts was developed by Yves Bontemps and is freely available on the open source software repository Sourceforge.38

Chapter 6 is based on two different case studies that both investigate with the same methodology the possibilities offered by user comments to augment the quality of metadata. In both cases, the content of a sample population of the comments was mapped with the help of the Shatford classification to the content of a sample population of user queries, in order to identify the utility of the user comments. In both cases, a confidence interval of 5 percent and a confidence level of 95 percent were applied to obtain the sample populations. The first case study of this chapter consists of user comments from the image database of the National Archives of the Netherlands, and offers a huge amount of historical photographs from the national press photo agency. The database appeals to the general public and the comments come from laymen and amateur historians. The Ross Archive of African Images offers the second case study, and differs from the first case study by its highly specific content and closed access to a limited number of domain experts. These maximum variation type of case studies allow me to compare the utility of two different types of comments.

The last case study, which figures in Chapter 7 introduces the idea of pro-actively following up the concrete metadata needs of users. In collaboration with the developer of an open source collection management system, a dynamic search- and display interface is designed. Users of the administrative back-end can add and delete metadata fields in a very in-

38 See http://sourceforge.net/projects/dataprofiler/.
tuitive manner to customize the search- and display interface according to
to their personal needs. The interaction between the users and the search in
interface delivers statistical information on how and which metadata fields
are being used by users and which are not. The prototype of this applica
tion is implemented within the collection registration database of the
September 11th museum in New York. The use of the prototype delivers
statistical information on how each user interacts with the set of availa
ble metadata fields during the testing period (November 2008 - January
2009). The September 11th museum is a critical case for the testing of
the prototype, as the institution and its holdings are both recent. The
museum has employed a team of catalogers who are currently actively
descriving the collection, which is very much a work in progress. New
types of objects, such as elements from the melted steel construction from
the World Trade Center or a crushed firetruck, are added to the collection.
This evolving character of the collection offers an ideal test bed for the
dynamic interface tool.

<table>
<thead>
<tr>
<th>Chapter 5</th>
<th>Organization</th>
<th>Type</th>
<th>Specificity</th>
</tr>
</thead>
</table>
|            | Royal Museum of   | Critical   | The volume and the heteroge
central Africa | neous character of the collec
tion management database pro|
|            | image database    | Maximum    |vide an excellent case for the|
|            |                   | variation  | data profiler |
|            | Section 6.4.2     | Case of    | Case of com
|            | National Archives | comments by |
museum of African       | laymen      |
|            | Images            | variation  | Case of comments by experts |
|            | Chapter 7         | Critical   | Recent and evolving nature of |
|            | September 11th    |            | the collection |
|            | museum            |            |             |

Table 1.3: Main case studies presented in the second part of the dissertation

1.4 Structure of the dissertation

Interdependence of the two parts

The thesis is built up of two parts, each part consisting of three chapters.
Two general outlines, available on pages 41 and 121, summarize the
content of the parts. Each individual chapter is preceded by a table of
1.4. STRUCTURE OF THE DISSERTATION

contents and a summary of the chapter. A general introduction in Chapter 1 and the final conclusions in Chapter 8 introduce and close up the dissertation.

Both parts mutually influence and depend on one another. The first part delivers an original and critical state of the art on the issue of metadata quality in the cultural heritage sector, by theorizing unacknowledged aspects of metadata quality and adopting a historical view on digitization practices. The second part brings in the necessary empirical research to back up the assertions made in the theoretical approach of the first part.

The two way interaction between the first and the second part is facilitated by the concept of the stratified time, presented in Section 1.2.1, which intervenes in and structures both parts. The theoretical first part mainly focuses on evolutions from the long and intermediate time, such as the commodification of cultural heritage and the different technological transitions of documentation practices. The three operational strategies of the second part focus on the day to day metadata creation and management practices which are analyzed in the short time.

However, as was demonstrated in Section 1.2 in which the stratified time concept was introduced, there exists a continuum between the three time levels. This continuum is also reflected in the interaction between the first and the second part. The three case studies from the second part will illustrate how the short time can have an impact upon the structure of a metadata schema, a matter which is discussed in Chapter 3.

Structure of the first part

After the general introduction, the first part presents the issue of metadata quality in the cultural heritage sector and the different research questions in three chapters. Chapter 2 analyzes the specificity of using ICT within the cultural heritage sector. The design, implementation and use of software can be problematic throughout all possible application domains. It is therefore crucial to identify and underline the uniqueness of the problems that appear when applying ICT within the context of memory institutions. This specificity is defined in three points: the unclear identification of user needs, the role of cultural heritage within society and the difficulty of describing connotative content. These three elements are responsible for a unique problem area in regards to metadata quality. Section 2.2 on the role of user needs and Section 2.3 on the role of cultural heritage within society both describe evolutions which are placed within the long term.

Chapter 3 provides an original historical overview of the use of ICT for the documentation of cultural heritage and offers an overview of the transitions within the intermediate time. The confrontation between the hopes and expectations surrounding the use of ICT in the cultural heritage sector of the 1960s and 1970s and those expressed today show in which fields
The overview of the origins and the evolution of metadata themselves investigate how metadata practices differ from traditional cataloging and indexing. A more nuanced interpretation of metadata is defended, instead of insisting on a clear separation with past practices. On the other hand, new technological tools clearly affect the content they produce. Especially the "black box" character of collection registration software can have a negative impact upon metadata quality, as metadata practitioners cannot always customize these tools specifically in function of their needs. Not only the content of the metadata is influenced by the tool, but the profile and the status of the work metadata practitioners is also affected by new technological developments. Chapter 3 then focuses on two recent developments which have both been hailed by their advocates as the future of metadata: ontologies and folksonomies. Here again, the real added value and feasibility is evaluated for the cultural heritage domain. The concept of metadata mash-ups and offering metadata in a use-neutral manner ends the chapter.

The existing research and work on metadata quality in and outside the specific application domain of the dissertation is presented in Chapter 4. Although there is a consensus on the importance of metadata, the effective evaluation of its quality is not evident. The corporate sector has in recent years realized that bad quality data are responsible for important financial losses. Some of the specific characteristics of the cultural heritage sector, such as its low critical and economical stakes, have prevented it until now from becoming an issue. Concretely, the chapter starts out with some examples from the biodiversity research illustrating techniques and tools mobilized to tackle the issue of metadata quality. Within the specific domain of cultural heritage, the large national libraries were the first to reflect upon metadata quality. In the context of copy cataloging, a discussion rose on the balance between adhering to strict cataloging rules and user needs/cataloging costs. This discussion has led to a more pragmatic approach, that spread out to other players within the heritage sector, where a quicker access is increasingly preferred above striving toward the highest possible quality of metadata.

**Structure of the second part**

The second part of the dissertation gives the research the necessary empirical foundation. The majority of the academic research on metadata makes abstraction of the day to day practices of metadata creation. The acronym YAMS, which stands for "Yet Another Metadata Standard", illustrates the growing cynical attitude of metadata practitioners on the proliferation of metadata schemes, of which development sometimes seems to be an end in itself. But despite the necessary empirical character of metadata research, existing research often lacks an understanding of the
1.4. STRUCTURE OF THE DISSERTATION

day-to-day constraints and limitations institutions and metadata practitioners have to deal with on the workfloor. The rapid evolution of descriptive standards and of the technology needed to implement new metadata standards stands for example completely opposite with the very slow pass at which changes can be implemented within cultural institutions. More importantly, almost all institutions are confronted with legacy metadata. As fundamental as the work and research on metadata schemes and standardization is, they can rarely be fully applied as a fresh starting point on the work floor. The three operational strategies for enhancing the quality of metadata presented with the second part fully acknowledge this practical imperative.

The three strategies are presented in a logical order. Chapter 5 delivers the necessary diagnostic information on the formal quality of metadata. The methodology of data-profiling, which allows to obtain statistical information on the formal quality of metadata, is tested with the help of an open source data-profiling tool on an export of the metadata records from the ethnographic collection of the Royal Museum of Central Africa. The five simple analyses contained within the data-profiler illustrate how strategic information on a large set of metadata can be obtained in a quick and straightforward manner.

After the analysis of the formal characteristics of metadata, Chapter 6 investigates how an institution can act upon bad quality metadata. The previous chapter only took the formal quality of metadata into consideration and could not express any opinion regarding the actual content of the metadata. As the manual correction of metadata by professionals is too expensive and not feasible, Chapter 6 proposes to investigate the possibilities offered by user comments to augment the quality of metadata within cultural heritage databases. Two maximum variation case studies deliver empirical data on the relevancy of user comments. The first case study consists of the image database of the National Archives of the Netherlands, which offers a huge amount of historical photographs that appeal to the general public. The Ross Archive of African Images offers the second case study, and differs from the first case study by its highly specific content and closed access to a limited number of domain experts. The mapping of the user queries and user comments will demonstrate a correlation between the content of both, which illustrates the relevance of the user comments.

Chapter 7 provides the final approach by introducing the idea of proactively monitoring the concrete metadata needs of users. In collaboration with the development team of an open source collection registration software, a prototype is developed which offers a dynamic search interface to be implemented within a collection registration database. Users of this tool can customize their search interface in a very intuitive manner to their personal needs. Most importantly, the interaction between the users and the search interface delivers statistical information on how and which
metadata fields are being used by users and which not. This information can then be used to guide future actions upon the metadata schema and influence the long-term metadata strategy of an institution. A prototype of the tool is implemented within the collection registration database of the September 11th Memorial and Museum.

A final chapter presents the overall conclusions of the dissertation, which focus on five key outcomes. Firstly, the results from the operational strategies are presented. The utopian belief in the power and abilities of digitization practices to disclose the content of our cultural heritage has been countered throughout the entire dissertation with examples of how technologies do not always have a positive impact upon the quality of metadata. With the help of the stratified time concept, the thesis delivers the first holistic approach of metadata quality in the cultural heritage sector. The use-centric metadata quality perspective that was adopted throughout the dissertation both accentuates the importance of user needs but also points out in an original manner to the problems of a too narrow vision on user needs. A list of operational recommendations and future research possibilities complete the conclusions.

1.5 Original findings and outcomes

The novelty of this research lies in its critical review of the impact of digitization projects on the quality of metadata within the cultural heritage sector. A limited number of publications have already appeared on the digitization of cultural heritage, but none of these publications offer a thorough critique of both the discourse around digitization and its outcomes. Again, I can refer here to the work “Theorizing digital cultural heritage. A critical discourse” [30], which does not contain the critical discourse it promises to deliver. Sections 1.3.1.4 and 3.5.1 contain examples of my comments upon this work.

A significant larger research community exists on the topic of metadata. In 2008 for example, two important monographs were published specifically on metadata. Marcia Lei Zeng and Jia Qin offer the first comprehensive overview of all the different research aspects of metadata, including metadata quality [163]. But the merit of the book lies in the extensive review of existing publications and its pedagogical qualities. It does not provide any fundamentally new views on metadata. The publication of Muriel Foulonneau and Jenn Riley [56] offers a very practical handbook for the management of metadata within digitization projects and large scale digital libraries. But both books do not acknowledge the continuum that exists between traditional indexing and cataloging and the use of metadata.

The following points provide an overview of the specific original findings of the dissertation:
1.5. ORIGINAL FINDINGS AND OUTCOMES

- **First critical review of a decade of digitization projects in the cultural heritage sector:** after a decade of experimentations with the digitization of cultural heritage, time has come to analyze the outcomes from this period. Both private and public bodies are now structurally investing considerable resources in the digitization of cultural heritage, but the issues of sustainability and digital preservation are far from being resolved. On a more general level, there is hardly no awareness on the potential dangers and the negative impact new technologies can have for the documentation of cultural heritage. Chapter 2 defines the specificity of the problems the cultural heritage sector is confronted with when implementing new technologies. The dissertation then fills in the lack of knowledge on possible negative side-effects of new technologies by focusing on the issue of metadata quality and how it is affected throughout different documentation tools in Chapter 3 and 4. The continuum I draw between traditional documentation practices and metadata management allows an original evaluation of the added-value of digitization practices.

- **Original framework for understanding metadata quality:** in order to acknowledge the continuum between traditional documentation practices and the current work on metadata, the concept of the “temporalités étagées” is developed within the introduction of the dissertation. Building upon the work of Braudel [27] and Boydens [25], I adapted this conceptual framework to my specific application domain. It is the first framework that combines the impact of 1) long term cultural shifts and evolutions in government policies, 2) the change in technology and standards and 3) the evolution of the metadata themselves upon metadata quality within the cultural heritage sector.

- **Three strategies for tackling metadata quality in practice:** the second part of the dissertation, consisting of Chapters 5, 6 and 7, bundles three original case studies that demonstrate strategies to approach metadata quality in practice. By adopting very pragmatic, low-cost and non-resource intensive tools and methodologies, the barriers for the uptake of these tools and methodologies by all types of heritage institutions are kept as low as possible. The case studies also demonstrate the strength of the stratified time concept, by illustrating how the daily reality of metadata creation and management of the short time can influence the intermediate and long time.
CHAPTER 1. INTRODUCTION
Part I

Understanding metadata quality: a critical and original state of the art
Outline of the first part

The dissertation is divided into two major parts. This first part sketches the necessary historical and technical background of metadata practices within the cultural heritage sector, which act as the point of departure for the development of concrete metadata quality strategies that will be presented within the second part of the dissertation.

This introduction to metadata practices is in itself divided into three chapters. I start out by identifying the specificity of cultural heritage when working with ICT in Chapter 2. The design, implementation and use of software can be problematic throughout all possible application domains. Therefore, it is crucial to identify and underline the uniqueness of the problems that appear when applying ICT within the context of memory institutions. This specificity is defined in four points: the unclear identification of user needs, the non-critical and - economical nature of heritage, the convergence of libraries, museums and archives and finally the difficulty of describing connotative content. Important developments from the “temps long”, such as the shifting view on the economical and social value of cultural heritage, are presented.

Chapter 3 provides the historical context of the use of ICT for the documentation of cultural heritage. The analysis of the work and research undertaken in the 1960s and 1970s allows a better understanding of the impact innovative technologies have had upon cataloging and indexing the last four decades. New tools engender new practices and also affect the profile of the users of these tools. Apart from the evolution of tools such as software, the change in documentation standards and practices also need to be examined. Both elements correspond to what I have identified as the “temps intermédiaire” that impacts metadata quality. Folksonomies and ontologies, two examples of recent documentary methods which both figure at the extreme ends of documentary vocabularies, are examined. The most recent approach, that focuses on offering metadata in a use-neutral manner for so called mash-ups concludes the chapter.

The existing research and work on metadata quality in and outside my specific application domain is presented in chapter 4, which wraps up the first part. Examples from the biodiversity research illustrate techniques and tools that are mobilized in other domains to tackle the issue of metadata quality. Within the cultural heritage sector, the large national libraries were the first to reflect upon metadata quality. In the context of copy cataloging, a discussion has risen on the balance between adhering to strict cataloging rules and user needs/cataloging costs on the other hand. This discussion has led to a more pragmatic approach, which spread out to other players within the heritage sector, where fast access is increasingly preferred on striving toward highest possible quality of metadata.
Chapter 2

Specificity of digitized cultural heritage

Contents

2.1 Introduction .................................................. 44
2.2 Unclear identification of the users and their needs . 44
  2.2.1 Existing research on needs assessment ........ 45
  2.2.2 Practical problems with user needs assessment . 47
  2.2.3 Conceptual problems with user needs assessment 48
2.3 Role of cultural heritage within society ............ 53
  2.3.1 Increasing the social relevancy of heritage by di-
       gitization ............................................. 54
  2.3.2 Increasing the economical relevancy of heritage
       by digitization ....................................... 55
2.4 Dealing with connotative content .................... 58
  2.4.1 Central role of images within digitization projects 61
  2.4.2 Specificity of images as documentation aid .... 63
2.5 Conclusions ................................................. 64

Summary: The implementation of new technologies entails problems throughout all application domains. However, the specific character of the cultural heritage sector is responsible for difficulties encountered within digitization projects. This chapter focuses on the three aspects of cultural heritage responsible for these difficulties: the problem of defining user needs of cultural heritage, the weak position of the cultural heritage sector within society and the connotative character of cultural heritage. The implications on metadata quality of these three characteristics are discussed within the conclusions.
CHAPTER 2. SPECIFICITY OF DIGITIZED CULTURAL HERITAGE

2.1 Introduction

The problems addressed within this dissertation are on the crossroad of different disciplines, practices and epistemological traditions. However, the central element within this research is the impact of ICT upon the documentation practices within the cultural heritage domain. As ICT has permeated every sphere of society, its impact is noticeable in all application domains. The problems that have risen while implementing technologies within the cultural sector are therefore not always unique. The work of Paul David states that the introduction of ICT tends to lower productivity, as many people on the work floor try to hold on to old working practices within the new digital environment [38]. The widely quoted “Chaos report” published by the Standish Group in 1994 for example disclosed the very limited number of ICT projects that can be considered successful: 31% are stopped before the end of the project, 52% exceed the estimated budget and planning while only offering a limited number of the project goals, and only 16% can be considered a success.1

Although the report and its result are criticized [63], the figures do mark a problematic trend and prove that the difficulties encountered with the implementation of new technologies are not a unique issue for museums, libraries or archives. However, the specific character of this sector has an impact on the way technologies are implemented for the documentation of heritage and also on why this can be problematic. This chapter presents these specificities and how they influence metadata quality in a direct or indirect manner.

2.2 Unclear identification of the users and their needs

Heritage institutions have a long standing problem with the definition of the exact needs of their users. Compared to other application domains, user needs regarding cultural heritage are mostly defined in very general and vague terms.2 This situation makes metadata evaluation difficult, since quality is at its most abstract level defined as the “fitness for purpose” (see Section 1.1.3.2). But how can the adequacy of metadata be judged without sufficient knowledge of user expectations?

The importance of this question can be underlined by using the metaphor of the construction of a tunnel. The creation and management of metadata on the one hand, and their use by the public on the other hand,

---

1See the article of Michel Volle on his website: http://www.volle.com/lectures/glass.htm.
2The report “Assessment of End-User Needs in IMLS-Funded Digitization Projects” [1], produced by the Institute of Museum and Library Studies in 2003, offers the best overview on how the heritage institutions struggle with the issue of user-needs in the context of digitization projects.
2.2. UNCLEAR IDENTIFICATION OF THE USERS AND THEIR NEEDS

can be considered as the two ends of where the construction of a tunnel starts. The goal is to meet in the middle and to have a perfectly coordinated match in order for the creation process of the tunnel to be a success. The same logic applies in regards to metadata creation and management processes and user needs. The more the two are in line with one another, the closer comes the goal of metadata quality.

The next sections will focus on the existing research regarding needs assessment in the cultural heritage sector, but will also point out important difficulties, which remain largely untreated within the literature. Before entering the topic of needs assessment in detail, it is however important to clarify our understanding of the term “user”, and how digitization has engendered a major shift within this understanding.

Before the advent of the Internet, the actual users of heritage collections were limited to the researchers of the institution which held the collection, and to a lesser extent external researchers and in the case of archives, genealogists and amateur historians. The limited number of users stood in relation with the restrictions in space and time to physically access the collection, although some surrogate use through the delivery of photographs and photocopies was possible.

Digitization projects have drastically altered this situation in the sense that the potential user base now also incorporates the general public which has an interest in discovering its own cultural and historical context.

2.2.1 Existing research on needs assessment

Libraries have the strongest tradition of performing user need assessments within the heritage sector, as their task is by nature more access-oriented. A large number of publications are available on needs assessment, both on libraries in a traditional, paper-based, and a digital environment. Some actual assessments are also available online, which give an overview of the methods used.

User needs are less well studied in the archival world. Sundqvist gives an overview of the different types of user studies that have been performed within the archival sector. Traditionally, a lot of attention is given to the definition of user groups, which are mostly split up over scholars, teachers, journalists and other professionals on the one hand (“vocational” users) and members of the larger public such as amateur genealogists (“avocational users”).

Although museums traditionally tend to let the preservation of their holdings prevail on their access, the general users have a growing belief that

---

3 See for example [98] and [158].

4 See for example the study performed at the University Library of Amsterdam [148] for a recent and detailed study.
in a digitized context, museums resources should be as easily available as
electronic resources in libraries [100]. The digitization of cultural con-
tent implies new communication and usage processes between museums
and their public. Content and objects no longer are solely exposed on
site within a museum, but mainly reach their audience through collection
databases, the website of the institution, or even on some mobile
device [132].

While there exists a strong will amongst the museum community to dis-
cover and define the evolving needs of their public, no clear understand-
ing about users and their expectations has been gained over the last few
years. As Darren Peacock notes, “this confusion results in research and
analysis in the form of inconsistently modeled snapshots of what is hap-
pening on-line. Despite the large amount of effort going into user research
over more than a decade, we still have a very fragmented understanding
of users and the user experience on museum Web sites” [115]. In his arti-
cle, Peacock identifies four different terms that have been used to refer to
the public that visit museums. Each term reflects a specific paradigm and
assumptions:

- **the audience**: this term refers to a subject, passive population, as
  identified within the context of traditional broadcast media such as
  the radio and television. This model has its roots in the 19th century
  patronalistic ideas about how cultural institutions could attract and
  educate the masses.

- **the visitor**: the public is considered as visitors that make an active
  choice to visit a museum. The visitor studies paradigm prefers an
  ethnographic methodology to study visitor “behavior”.

- **the consumer**: this approach is driven by market research that fo-
cuses on how needs can be satisfied and relies on “behaviorist models
  of stimulus and response”. As Peacock adequately notes, this “market
  conception completely alters the relationship of supply and demand
  for museums, inverting the dependency of consumer on producer to
  its reverse.”

- **the user**: usability analysis focuses on the way agents can achieve
  their tasks within an interface. User-centered design tests rely on
  experts or representatives of users to test the fitness for use of search
  and display interfaces.

Section 2.2.3.1 will elaborate on the consumer paradigm and Section
2.2.3.2 on the difficulties of identifying user needs within the cultural
heritage domain.
2.2.2 Practical problems with user needs assessment

Performing needs assessment can be very tedious in practice. One of the most interesting recent projects that can illustrate some of the difficulties encountered on the terrain is the Museums and the Online Archive of California (MOAC). At the end of this project, a new two year project was run to evaluate the work done to increase and enhance access to museum, library and archival collections from California. In spite of the large scale funding and the publicity the project attracted, the final results of the user evaluation project are rather disappointing. The unpublished report points out the difficulties that were noted during the project [61]. The interest of this case is that the mentioned difficulties can be generalized to other projects.

The researchers counted on using pre-existing use data from the project partners, but these where almost non-existent amongst the museum partners. Only the libraries had statistical and qualitative data on the use of their collections. Differences in willingness to perform usage study were thought to be at the basis of this differences between libraries and museums. Libraries mostly have one single point of contact such as the reference desk, whereas the museum user is thought to contact multiple places and persons, such as the curatorial, registration, education and public relations department [61].

The analysis of transaction logs had to reveal how many and from which domain users were coming, but due to “institutional challenges involved in identifying resources to support specific reinstrumentation for the research project” and some technical problems, the log files could not be used. However, web traffic analysis is far more complex than it seems on the first hand. The most frequently cited parameter to express the online success of a website is the number of client requests made to the web server, or commonly known as hits. Different methodologies, such as logfile analysis versus page tagging for example, exist which result in different outcomes. The ongoing technological evolutions regarding the use of web caches, the indexing of webcontent by search engine spiders and dynamic IP addresses have made it increasingly difficult to identify a unique human visit of a webpage. A popular way to boost the number of hits of a website is to publish the whole content of a database (which can contain hundreds of thousands of records) on the Internet, where the content is then massively indexed by search engines. This results then in an artificially augmented number of visits. So when analyzing web usage statistics, it is crucial to validate the number of hits with the overall character of individual user sessions. Nevertheless, the number of hits of a website does offer crucial information on trends and the evolution of the use of a website over a longer period of time.

\[\text{The following information regarding the analysis of web traffic is based upon [95] and [94].}\]
Another traditional way of collecting information from users are web forms. In the context of the MOAC-project, only 29 persons actually completed the form between September 2003 and January 2006. Detailed questionnaires were also made for each user group targeted in MOAC II (K-12 teachers, undergraduates, faculty, and information professionals). A lot of the participants backed out of the interactive web form before finishing up all of the questions. Lastly, some “Think-aloud” sessions were organized, where selected participants (teachers, faculty, and information professionals) in Los Angeles and San Francisco were asked to complete two tasks using MOAC: a task previously predetermined by the research team and an original task of the participant. With the use of specific software, the participants’ key strokes were tracked and their voices and images of them were recorded for further analysis. Each session supposed to take an hour. Despite serious recruitment efforts, only few people were found to participate. The project exactly illustrates the difficulties to obtain sound results from a user survey and focus groups.

2.2.3 Conceptual problems with user needs assessment

In Section 2.2 the metaphor of a tunnel was used to explain the importance of user needs, to insist on the fact that for metadata to be of good quality, the content of the metadata should meet the requirements of its users, as both ends of a tunnel should eventually meet somewhere in the middle. However, when playing the devil’s advocate, one could enumerate countless examples from our cultural heritage canon which illustrate the dangers and the fugacity of the mainstream appraisal of cultural heritage. Vincent van Gogh famously only sold one painting in the last year of his life, whereas his museum is now one of the main touristic attractions of Amsterdam.

Throughout the literature on metadata quality, there is a consensus that the satisfaction of user needs is the central element when evaluating quality.6 The importance that is attributed to user needs within the current digitization discourse can be identified as one of the key elements amongst the long term evolutions that influence metadata quality. The following sections also illustrate the impact of the inherent changing nature of user needs. Chapter 7 will propose in the second part of the dissertation an operational strategy to monitor in permanence user needs, as a way of confronting this problem.

2.2.3.1 The commodification of cultural heritage

Among the four paradigms defined by Darren Peacock in Section 2.2.1, the “consumer”-model has had the biggest impact during the last decade.

---

6See for example [25, 28].
The cultural heritage sector has to position itself more and more on the common marketplace. The fact that webshops are becoming an unexceptional feature of cultural heritage websites illustrates this tendency. In 2005 the Flemish government invested considerably in the Reproductiefonds.\footnote{See the website of the organization on http://www.reproductiefonds.be/} This organization centralizes the high-resolution images of the major heritage institutions in Flanders in a database, from which they can be downloaded after payment. Derived products such as mugs, puzzles, illustrated with the masterpieces of their collection can also be purchased online. Figure 2.1 illustrates the types of merchandising that are offered. The rhetoric of the consumer has also effected the role of the cultural heritage professionals, which are now sometimes referred to as “information brokers” [30].

Figure 2.1: Screenshot of the webshop of the art image site Lukasweb, which distributed high-quality digital reproductions of Flemish art

Cultural heritage taken hostage by new technologies

The commodification of cultural heritage is one of the important evolutions that can be placed within the “temps long” of the stratified time concept. Policy makers from public administrations at European or national levels have played a major role in the uptake of the discourse on
the commercialization of cultural heritage. One of the most illustrative reports that document this discourse is the report “Technological Landscapes for Tomorrow’s Cultural Economy: Unlocking the Value of Cultural Heritage”, published under the DigiCult programme of the European Commission [59]. This document reflects the European Commission’s view on digital cultural heritage in a very explicit way.

In the DigiCult-report, the heritage sector is seen as an important content provider for the ICT industry:

> What technologists have come to understand is that technologies are not taken up if content is inadequate (hence the rise of texting and the failure of WAP). Future online services require a critical mass of quality content, and the proven popularity of museums on the Web suggests a natural way forward. Certainly there is a relationship of mutual benefit here between those who wish to develop new technological solutions and the content-holding institutions who wish to develop new audiences. And as cultural materials in museums, libraries and archives are traditionally available free, they also present a relatively unproblematic resource from which to tap. This relationship has given the DigiCULT initiative a certain prestige and political weight [84, p. 136].

This quote demonstrates the sometimes problematic relationship between technologies and cultural heritage within the context of digitization projects. By financing digitization projects that have public sector and private partners, governments at different levels have tried to, as the title of the DigiCult adequately states, to “unlock the value of cultural heritage”. In practice, many projects which are centered around a new type of technology, such as hand-held and mobile devices, reduce cultural heritage institutions as easy content-providers of free content for a technology that needs to get launched on the consumer market.

**The dangers of placing cultural heritage as a regular player on the market**

When interpreted literally, the “fitness for purpose” metadata quality definition as presented in Section 1.1.3.2 refers strongly to the idea of self-regulating markets, where the demand has a direct impact on the offer. Francois Matarasso, a consultant on the evaluation of the arts, wrote an article on the difficulties and dangers of applying the “fitness for purpose” definition in a too narrow or literal manner within the cultural heritage domain. He explains that free market regulation cannot be the only quality criterion:
The market is never as free as we might wish it to be. There is a strong commercial interest in a standardization of taste, whether in music, films or coffee, since it enlarges markets and reduces overheads. These forces are dangerous enough when it comes to ordinary commodities: they are all the more when it comes to the space (= the cultural sector) where we shape, question and transit our values. (...) The views of audiences and participants in the arts are a component of evaluation. Only in the commercial sector do they act as a determinant measure of worth: and the point of having a public sector is precisely to introduce other values and safeguards to our cultural life” [103, p. 3]

The “standardization of taste” Matarasso refers to in his essay can be easily transposed to the very current and necessary debate that Paul Staincliffe called “Rules versus Access” that is taking place in the library community, but which is also affecting museum and archives practices and policies. Basically, Staincliffe criticizes the “Google generation” for obliging information professionals, such as library catalogers, to lower their thresholds, which negatively affects the quality of cataloging:

The “Rules versus Access” centers on the conflict between the use of standards and the provision of access. I am not implying that the two are mutually exclusive: merely that where once it was the rules that dictated how a resource was described, now it is the demands of those wanting access to it. Instead of practicing an Art or a Science, we have allowed resource description to be dragged down to the customers level. It has been “popularized” to an extent that negatively affects retrieval [133, p. 1].

Here again, we find an increasing tension between user needs and the authority of a heritage institution. Benjamin Barber has been reacting strongly against the privatization and commercialisation of the public sphere in general. In an essay written for Culturele Biografie Vlaanderen he reflects on this growing tension between the need to take into account the needs of the public and the preservation of cultural heritage:

We can argue endlessly about whether museums should be temples of preservation or public squares of democracy, but if in reality they are transformed into pop-cultural malls that commodify culture, both sides of the old debate will be losers, no matter which side ‘wins’. ... The general presupposition of the democratic context is that democratization of cultural preservation processes requires greater public deliberation and input into preservation - which however puts it into tension with traditional curatorial understandings of conservation and preser-
vation. While this narrow construction of democracy does create new democratic demands as well as tensions with curatorial and museum approaches, there is a deeper sense of democracy that may help mitigate the tensions and that suggest realistic strategies of heritage management suitable to democratic times [8, p. 83].

Barber then elaborates within the essay a broader view of democracy, by also taking into account previous and future generations. By referring to the concept of the “social pact” of Edmund Burke, Barber points out that democracy is a partnership not only between the members of the current generation, but also those who are dead, and those who will come after us. He writes that “commodification of culture is the tyranny of the present over past and future alike” [8]. In his essay, Barber refers to the demolition of the Buddha images in Afghanistan by the Taliban, as an aggressive act against the generation that build the monuments, but also toward future generations.

2.2.3.2 Providing in evolving and future needs

The previous sections illustrated the problem of interpreting the “fitness for purpose” metadata quality definition in a too narrow or literal sense. The broadened vision on the user, by including also past and future users, of Benjamin Barber brings us to one of the most important characteristics of user needs in the cultural heritage sector: the envisioned, the actual and especially the potential uses can largely differ, as cultural heritage institutions also address needs that do not exist yet. The impossibility to predict future uses is a long-standing problem within archival appraisal. The evolution of user needs clearly makes the selection process of what is included and what not within the holdings of heritage institutions a very tedious task.

We can go back to the very starting point of the professionalization of heritage institutions in the middle of the 19th century for an interesting Belgian example. One of the pillars of metadata quality within contemporary archival sciences is the notion of the “respect des fonds”, which refers to the crucial importance of keeping all of the documents in their original context. The importance of a single document is considered to be found in its place within the other documents from the same “fonds”, and not so much in the content of the document itself. This was not always the case. Louis Prosper Gachard, who was the first head archivist of the newly founded national archives of Belgium between 1831-1885, arranged archival items according to their content or formal characteristics.

---

8This logic is also presented in the following two reports: [112, 1].
9The following articles give a good overview of the most recent developments regarding archival appraisal: [49, 45, 72].
2.3. ROLE OF CULTURAL HERITAGE WITHIN SOCIETY

Paper documents were for example separated from parchment documents, or arranged in chronological order. This way of organizing and describing the archival holdings needed to facilitate the work on the historical legitimation of the freshly founded Belgian state [3].

This specific example illustrates once more the highly evolving nature of what is to be considered good quality metadata. Chapter 7 will therefore propose in the second part of the dissertation an operational strategy to monitor in permanence user needs, so that changes in needs can be detected as soon as possible.

2.3 Role of cultural heritage within society

This section will review how the digitization of cultural heritage has affected the vision of public funding agencies on the role cultural heritage can play within society. Both phenomena are a part of the evolutions from the “temps long”. In comparison with information systems from other domains, such as the financial or the administrative sectors, the economical value of metadata from the cultural heritage sector is comparatively limited. This clearly has an impact on how effectively metadata quality is followed up and taken seriously. As Kenneth Lindsay already stated at the conference “Computers and their potential applications in museums” in 1968:

> We have to realize that society - which must pay for expensive computer systems - may not share our concerns. Our enterprises do not have the earning power of the industry, the status of science, or the drama of defense.” [92, p. 44]

In her work on the quality of social security databases, Isabelle Boydens notes for example that the social security administration employed at the end of the 1990s around 300 persons to correct the 500,000 to 1,100,000 errors from the social security databases. These errors only represented between 0.05 and 0.09% of the overall values [25]. The financial and social stakes of the social security are so tremendously high, that every detected error needs to be corrected.

Apart from important economical interests, critical domains such as the medical sector can not accept errors within their information systems. In some extreme cases, people have actually died as a result of bad quality metadata. In 2001, a healthy 24-year old woman, who participated in a medical project at the John Hopkins university, received a lethal dose of hexamethonium. The reason behind this tragedy was to be found in the fact that the publications that pointed out the dangerous effects of the substance were not retrieved during the literature study done in the PubMed database by the researchers that were responsible for the project [2].
The situation is completely different within the cultural sector, where the economical and social stakes are harder to define. The level of detail of the documentation of valuable works of art or cultural heritage items plays an important role for insurance issues and for the detection of forgeries. But both of these situations only concern a small subset of the objects that are collected by cultural heritage institutions. Museums do collect unique, valuable items, but libraries and archives mainly retain documents that have no direct financial value on their own.

The debate on the utility of cultural heritage for society is not a matter to be addressed within this dissertation. The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines the interest of cultural heritage mainly in the areas of social and economical development. The next sections will analyze what the impact is of digitization of cultural heritage upon these two issues.

### 2.3.1 Increasing the social relevancy of heritage by digitization

Digitization projects have the ambition to augment the immediate outcome and importance of cultural heritage within society by attracting a larger audience. Digitization does not only offers the opportunity to bring heritage to a broader public, it also provides possibilities to incorporate new forms of heritage that did not make part of the official cultural heritage canon.

This widened scope on cultural heritage played an important part in levering the relevancy of heritage toward certain groups of society. A typical example is the community-driven website “Heuvelland Verbeeldt”, which offers the elderly of a small region in Flanders the opportunity to scan their collection of historical postcards or photographs and to describe and publish them within a database. The direct contact between the citizen and cultural heritage is considered a much valued aspect of digitization projects. Figure 2.2 illustrates the website, with the different user comments that are associated with a specific historical photograph. Chapter 6 will go deeper into the topic of user comments and on the tension between “functional” and “factual” memory these practices bring with them.

The evaluation of the social impact of cultural heritage through digitization projects remains problematic. Therefore, I will focus now mainly on the possibilities digitization offers to augment the economical value of cultural heritage.

---

10 See the UNESCO website for the definition of the goal of museums: “As points of access for knowledge about cultures and centers of education, museums also contribute to mutual understanding and social cohesion as well as economic and human development.” Quoted from [http://portal.unesco.org/culture/](http://portal.unesco.org/culture/)

11 The following publications focus on this issue: [15, 60, 30].

12 See the website of this project on [http://www.heuvellandverbeeldt.be](http://www.heuvellandverbeeldt.be).
2.3. ROLE OF CULTURAL HERITAGE WITHIN SOCIETY

Figure 2.2: Screenshot illustrating user comments sent to the local community-driven website “Heuvelland Verbeeldt”

2.3.2 Increasing the economical relevancy of heritage by digitization

Digitization projects have also tried to raise the economical value of heritage. In the early days of digitization, heritage institutions hoped that the digitization of their collections would generate new revenues [130]. Section 2.2.3.1, which focused on the commodification of cultural heritage, demonstrated the eager to develop commercial services based upon digitized heritage. Metadata could play a crucial role in the re-use and marketing of digital cultural heritage. However, American, European and other national reports and projects investigating business models based on the commercialization of digital cultural heritage from the public domain do not point to viable options or business plans.\(^\text{13}\)

Even the self-sustainability of digitization projects can in most cases not be attained. A 2008 report published by OCLC Programs and Research offers insights and results of the RLG’s Cultural Materials and Trove.net project, which were initiated out in 1999 and launched in 2002 with the specific goal to study the possibilities for collaborative large scale digitization projects to sustain themselves [52]. 54 international institutions participated and represented 117 digitized special collections, which included all to-

\(^{13}\)The European Visual Archives Market validation Project (EVAMP), which was funded by the eTen program of the European Commission, had the specific task of investigating viable business models for the commercialization of digitized historical photographs. The project website is unfortunately no longer available, but the project never was able to deliver a sound business plan.
gethеr almost a million digital objects. The goal of Trove.net was to make digitization self-sustaining for the participating institutions, but the different options that were taken to financially valorize the digitized collections were not successful. A corporate sponsor for the website was searched but not found. The inclusion of advertisement on the website through the use of AdWords and Yahoo! Sponsored Search words costed more than they returned. AdSense adds brought in less then 2000 dollar a year. Licensing the images through collaboration with stock agencies seemed to be the most interesting option. In the three and half years of agreement with a commercial stock agency, only 986 images were licenced which resulted in a total amount of 26,810.57 dollar. This makes an average amount of 27.19 dollar per image. A bit more then half of the contributors to Grove.net received some licensing fees, ranging from a couple of dollars to over a thousand.

When taking into account the overall effort that was put into the project, this is a very poor financial outcome. The reason behind the lack of interest for image licensing was the lack of interesting content for the corporate world. Big cultural heritage institutions may posses a limited number of very attractive and iconic images, that can be licensed at competitive rates, but the bulk of historical images is not of sufficient interest to build a viable business plan.

2.3.2.1 Customer-driven indexation: the example of Magnum

After the unsuccessful examples from the public cultural heritage sector, it is interesting to see how private players on the market manage to survive, and more importantly, to study whether it affects metadata quality. The prestigious press photo agency Magnum is one of the most interesting examples to study, as their database offers a very high level of content, and they are self-sustainable.\(^\text{14}\) Concrete figures of their financial turnover are not publicly available, but as Magnum receives no public funding whatever, they are obliged to be self-sustaining.

Magnum draws its biggest income from image licensing to the media and the corporate world. Their prestigious background and tradition of excellence in the photographic domain clearly distinguish them from the other important players within the stock agencies market, and allows them to charge considerable prices for the quality they offer.

The needs of their users, who consists of the written press and the corporate advertisement sector, have influenced the way Magnum documents and digitizes its collection. Their thesaurus, which has been developed in house, is peculiar. The most interesting feature of Magnum's thesaurus is the use of emotional and abstract terms, such as anger, love, tenderness, aggression, etc. Figure 2.4 illustrate the different abstract and emotional

\(^\text{14}\)See the agency's website on http://www.magnumphotos.com/.
2.3. ROLE OF CULTURAL HERITAGE WITHIN SOCIETY

terms Magnum has created within its thesaurus. The inclusion and use of these types of terms is rather uncommon, but reflects the needs of their user base, who want to find images from the Magnum archives to illustrate advertisements or news articles.

The thesaurus is maintained by the cataloging teams in Paris, London, New York and Tokyo, and different practices with the thesaurus reflect some cultural differences. The London bureau for example applies a lot more terms in relation to the social class of the persons that are depicted. The New York agency apparently applies a lot more terms on the physical appearance, and uses a lot more the term “obese” than their French colleagues for example.

Apart from these cultural differences, the thesaurus also reflects the impact users have had on the descriptive practices of the agency. In the past, the racial origin of portrayed people was only reflected within the descriptions if it made sense within the context of the image, if for example a European was depicted in a typical African scenery or vice versa. On the explicit demand of the customers, this policy has been changed and now all of the images with people have thesaurus terms attached to them which describe the racial origin of the depicted people. Figure 2.5 from the administrative back-end of the Magnum database where an image of
CHAPTER 2. SPECIFICITY OF DIGITIZED CULTURAL HERITAGE

Figure 2.4: Illustration of the different abstract and emotional terms Magnum has created within its thesaurus

Ian Parr is described, illustrates this.

So the case of the Magnum database clearly shows that not only the content itself needs to be of a very high quality, but the metadata also has to correspond to the needs of the commercial user base.

2.4 Dealing with connotative content

When considering a cultural heritage object as an informational resource, one of its specific characteristics is the importance I attribute to the connotative meaning of the resource. The distinction between denotation and connotation was made explicit by Roland Barthes [9]. The distinction between denotation, which is the literal or the primary meaning of a resource, and connotation, which corresponds to the suggestive or secondary meaning, has fundamental implications for information retrieval and metadata. Full-text searching for example is based upon the denotative meaning of resources, in the sense that search engines retrieve re-
2.4. DEALING WITH CONNOTATIVE CONTENT

Figure 2.5: Illustration of the use of racial origin descriptors ("type humain blanc") within the indexing process at Magnum

sults by matching the search string with the actual content of a textual resources. The same logic can be applied to CBIR software that detects forms, patterns and colors within the image itself. The connotative meaning of a resource on the other hand is made explicit by human indexing.15

The example Barthes used to illustrate the duality between denotative and connotative meaning has since then acquired an iconic status. The Panzani commercial, illustrated in Figure 2.6, represents on a denotative level a bag with groceries such as tomatoes, peppers and pasta. The primal colors of the image are red and green. But the meaning of the image does not reside within the objects that are depicted, but in the information that is associated with them, which is Italy and fresh food. These are the connotative meanings of the image.

15Progress is being made to automatically attach connotative meaning to digital resources with the help of ontologies, who could use the links between denotational semantics to automatically infer connotative information about the resource. But no functional prototypes are currently available in my application domain.
The value of cultural heritage resources resides particularly in their connotative meaning. It is the complex set of historically grown meanings which surround a cultural heritage resource that makes up its relevancy. One of the emblematic examples is perhaps the Black Square, which was painted by Kazimir Malevich in 1913 (see Figure 2.7). On a denotative level, we are confronted with a large black square, surrounded with a white border. But on a connotative level, Malevich himself interpreted the options to see either a black square resting on a white ground or a black hole surrounded by white sides, as pointers to the static facade and the inner dynamic of each object. Ever since, the painting has received a multitude of widely differing meanings and interpretations, which all depend on the contemporary context within which the meaning was formulated [37].

The example of the Malevich’s painting demonstrates that the value of cultural heritage lies in its ability to convey a relevant meaning throughout different contexts and periods. Chapter 6 will demonstrate how the users of online cultural heritage express their way of incorporating cultural heritage within their current lives, by helping the documentation process of historical photographs.
2.4. DEALING WITH CONNOTATIVE CONTENT

Although this tension between denotative and connotative meaning can also appear within other types of information resources, the importance of the connotative meaning is specifically present within iconographic resources. These resources have come to play a fundamental role within digitization projects. Apart from sound and moving image resources, all other analogue resources such as textual documents and other two or three dimensional iconographic resources are always converted into image files within the digitization process. Once textual documents are scanned, the resulting image files can, depending on the quality of the analogue original, be converted with the help of OCR techniques into a
digital textual document. The typology of documents in Section 2.4.2 will elaborate on the specificity of images.

Lithography at the end of the 18th century and photography in the middle of the 19th century provided the first means to reproduce images on a large scale. But after the second world war the military and the spatial industries, with their need to automate the creation and manipulation of large quantities of computerized images, laid down the foundations for the digital imaging. At the end of the 1980’s, desktop computers became more powerful and less expensive, which made them also more accessible for non-scientific or non-industrial domains, such as graphic design. VGA color displays replaced the monochrome ones, and the first Charge Coupled Devices (CCD) made their way to the consumer market in video cameras.\(^{16}\)

However, data storage and transfer remained problematic until the 1990’s, when cheaper and faster hard drives appeared. These different technological evolutions culminated toward 2000 to the point where both professional and amateur analogue photography was superseded by its digital counterpart. Ever since, a continuous mass of digital images has flooded every possible domain. Puglia [120] has defined four trends in digital imaging:

- from lower minimal spatial resolution to higher spatial resolution

- from 1-bit scanning, to grayscale scanning, and finally to color scanning

- from low-bit (8-bits per channel) to high-bit (16-bits per channel) for grayscale and color images

- from scanning for a specific purpose to digitizing in a “use neutral” manner

Even if the technology itself has become mature and easily available, there are very few professionnals within the heritage industry that really grasp the power of digital imaging. Taking into account the different analogue copies that already exist in many institutions of the original object, it is quite difficult to make a choice of which medium to digitize. Depending on the available budget and the quality needs of the resulting digital files, project managers need to decide whether to digitize the negative, the contact sheet or one of the different photographic prints that were created from a 19th century glass negative.

Low-priced high-resolution camera’s and scanning equipement launched a Do It Yourself-movement in the cultural heritage sector. The apparent straightforwardness of current consumer digitization equipement falsely

\(^{16}\)The work “The reconfigured eye: visual truths in the post-photographic era” of William Mitchell gives a good introduction into the history of digital imaging [105].
2.4. DEALING WITH CONNOTATIVE CONTENT

gives the impression that digital imaging requires no specific competencies.\footnote{A program manager of the Andrew Mellon foundation, which is one of the most important funders of digitization initiatives within the cultural heritage sector in the United States, launched in 2008 the idea on the Museum and Computers Network (MCN) mailing list to develop a “one button” digitization device, which could be handled by anyone without any experience in digital imaging. This illustrates the false belief that no specific expertise is required to digitize cultural heritage.} In incompetent hands, digitization worsens the spreading of corrupted copies, which was also the case with the invention of print, as Elisabeth Eisenstein notes:

“After comparing copies of a given reference work with early printed versions it is often found that an age-old process of corruption was aggravated and accelerated after print. \(\) But in the very course of accelerating a process of corruption, which had gone on in a much slower and a much irregular fashion under the aegis of scribes, the new medium made this process more visible to learned men and offered a way of overcoming it for the first time. \(\) In the hands of ignorant printers driving to make quick profits, data tended to get garbled at an even more rapid pace.” \cite[47, p. 82]{Eisenstein1979}

2.4.2 Specificity of images as documentation aid

Traditionally, the catalog or collection management system is built up around the metadata records. With the rise of digital imaging, the digital assets, which are in many cases images, have become the central element around which the metadata are organized \cite{Zeller1991}. As photographs came to play an important role as a documentation aid and considerable image collections were built up in museums and documentation centers, people became aware of the difficulty of providing an intellectual access to image collections. In both an analog and digital context, the description of images is seen as a complex and time consuming activity. Why exactly are images considered to be so much harder to describe than texts?\footnote{The work “Image retrieval. Theory and research” of Corinne Jorgensen provides an excellent state of the art on the problem of image indexing and retrieval \cite{Jorgensen2002}.} What is their specific position in the larger family of digital documents? All too often, these questions remain unanswered in the current literature on image metadata. Placing the digital image within the larger typology of digital documents allows us to define more clearly the borders with other types of documents. I rely on the typology of Zeller \cite{Zeller1991}, that proposes the following characteristics:

- Document = type of data + the type of form (encoding, structuration) + type of metadata \footnote{In his typology, Zeller proposes an extra category, the type of form of validation. I omitted this category as I do not find it has a discriminatory effect between types of documents.}
CHAPTER 2. SPECIFICITY OF DIGITIZED CULTURAL HERITAGE

- Type of data: Digital images are non-structured data, as opposed to structured data like for example database entries for a date field, that can only contain eight numerical values, the first two indicate the day, the third and fourth the month and the last four the year. In general, non-structured data are referred to as documentary data.

- Type of form of encoding: Digital images are documents based on graphical code (pixels), whereas digital textual documents are based on a finite number of alphanumerical characters of the ASCII alphabet. The ASCII characters in text files form the correspondence between the digital bit patterns and the symbols of written language, which ensures their automatic processing. Search and retrieval actions within textual documents are based on the recognition of alphanumerical characters by their bit sequences in the ASCII alphabet. Digital images on the contrary do not possess a finite character set for encoding the information contained in the file.

- Type of form of structure: The strong formal character of language facilitates computational operations on textual documents, whereas the internal organization of information in images has a notably less formal character [83]. No predefined grammar exists for the organization of images. But at a certain level, objects or physical features can be automatically retrieved, as illustrated by Content-Based Image Retrieval (CBIR) techniques.

- Type of metadata: Compared to textual documents, images offer very little intrinsic metadata, such as author, date and place of publication, for the simple reason that most images are unique and non-published documents. Header information in some image storage formats offer the possibility to create metadata such as the date of creation and technical specifications about the technologies used for the image production. But the large scale storage of this type of files is highly problematic. As for extrinsic metadata, several types of indexing languages have been developed specifically for images.

In summary, digital images can be considered as non-structured data, based on a graphical encoding with an informally structured content, with few intrinsic metadata. One aspect that is not represented within this typology is the conflict between generic versus specific and denotative versus connotative meaning, which is typical for visual documents. Images always represent a specific instance of something, in contrast to textual documents that can remain on a generic level [156].

2.5 Conclusions

The previous points of this chapter have identified the specific characteristics of cultural heritage which negatively affect metadata quality. The
three following elements are linked here with the specific impact they have upon metadata quality.

**Unclear user needs cause inadequate metadata schemes**

The difficulties of defining user needs have a clear impact on metadata quality. User needs are normally the starting point when defining the different metadata fields for a certain type of resource. When these needs are unclear or evolving throughout time, the metadata schema that is drafted will not fit the requirements.

The example of the evolution of the museum catalog of the city museum of Tienen and the resulting metadata problems, which are discussed in Section 1.2.2, illustrate these problems. When the first database for the management of the collection was created in FileMaker Pro, the metadata fields that figured on the paper-based cataloging cards preceding the database, were exactly re-created within the FileMaker Pro database. However, these fields clearly did no longer met the needs for the description of the collections as there occurs a huge number of fields that are never used and other fields which contain different types of metadata that should be split up over different metadata fields.

In order to counter the problem of ill-defined user needs on the behalf of the collection holder, I have developed a dynamic interface tool which will be extensively presented in Chapter 7. This tool offers the opportunity to analyze how the current metadata schema of a collection meets the needs of the catalogers.

**The non-crucial position of cultural heritage leads to short-term and non-structural funding**

Section 2.3 has focused on the impact of cultural heritage within society on a social and economical level. Compared to other application domains, such as the financial or medical sector, the importance of cultural heritage toward society is less direct. Digitization projects claim to enhance the social impact of cultural heritage and to offer new possibilities toward commercialization. Examples showed that new user groups now interact in alternative ways with cultural heritage. However, the initial expectations on the commercial possibilities of digitized cultural heritage seem for the moment largely unrealistic.

The limited funding within the cultural heritage sector has led to the situation where the digitization of holdings is mostly financed by external funding, which is limited in time. At the end of the 1990s, digitization was still not seen as a core task of a cultural heritage institution, but more as an experiment and a supplementary and secondary service. This situation
has fundamentally shifted within the last five to ten years, as digitization is now considered one of the top priorities of heritage institutions. Unfortunately, the financing of digitization is still largely based upon external and short-term funding. It is somehow ironic that Bishoff and Allen refer in their report “Business Planning for Cultural Heritage Institutions” to possibilities offered by sponsorship, advertising, foundations and donors within their chapter on “Planning sustainability” [17].

Governments at all levels are investing considerable resources. The European Commission for example has since 2000 invested each year 120 million euros for the development of Europeana [65]. But these external investments are made in the framework of specific programs, which aim to present visible results after a period of two or three years. This model of financing digitization projects has a perverse effect, in the sense that it prevents institutions of developing their own individual long-term digitization strategy. As funding can only be obtained by following the agenda of digitization programs, cultural heritage institutions are currently obliged to adapt both the content and the form of their digitization projects to accord with the reigning paradigms of policy makers. This situation has a negative impact upon metadata quality, albeit in an indirect manner. The current hype on user-generated metadata and content has resulted in a large number of funded projects that implement web2.0 tools and interfaces. Although user-generated metadata can be useful within a specific context (see Chapter 6), these projects take up resources that could be attributed to the creation and maintenance of professionally created metadata. However, it is rather difficult for policy makers to achieve appealing and short term results by investing in professionally created metadata.20

The connotative aspect of cultural heritage as a bottleneck for metadata creation

The advances made in the field of digital imaging have made it technically possible to digitize cultural heritage at both a very high qualitative and quantitative level. The current bottleneck for the success of digitization projects lies in the costs for the description of the created digital images. There is a clear consensus that the creation of metadata in general, and especially in the case of images, is a very resource-intensive activity.21

As the nature and the content of collections widely differ, it is not feasible to develop clear estimations for the cost of metadata creation, based upon averages from other projects. In the case of repurposing existing

---

20 Geoffrey Bowker notes a similar situation in the natural sciences, where it is notably easier to obtain funding for “charismatic” research on the panda bear then to study seaweed[25].

21 Poullommeau and Riley provide an overview of the different studies made on the financial implications of metadata creation and management [56].
metadata, project managers need to preview resources for adapting the existing metadata.\footnote{See the section on metadata reuse in the book of Zeng and Qin\cite[pp. 276-277]{Zeng}.} Multiple examples within the dissertation, such as the metadata from the city museum of Tienen (see Section 1.2.2) that illustrated the different emanations of the museum’s catalog throughout its history, and the metadata of the Royal Museum of Central Africa upon which data-profiling techniques will be applied in Section 5.4 illustrate that the migration and reuse of metadata often leads to metadata quality problems.

The importance that users now attribute to the presence of digital surrogates of resources in the form of images, and the complexity of presenting good quality metadata with the images has finally resulted in a situation of interdependence between metadata and images. Collection registration databases increasingly only publish the metadata record on the front-end website of an institution if the metadata record in question is linked with a digital reproduction of the resource. On the other hand, the existence of metadata is also increasingly used as a criterion for digitization\cite{Zeng}. If no metadata are available, then digitization project managers increasingly find it dangerous to start with the digitization of a collection, as they realize that the creation from scratch of metadata can form a serious threat to the success of a project.
Chapter 3

Applying ICT for metadata creation: an overview

Contents

3.1 Introduction ............................................. 70
3.2 Mismatched expectations .............................. 70
  3.2.1 Overview of the evolutions ....................... 70
  3.2.2 Confronting expectations with reality ............ 74
3.3 Metadata: old wine in new bags? .................... 76
  3.3.1 Overview of the evolution of metadata .......... 77
  3.3.2 Specific metadata schemes for the cultural heritage domain .................................. 81
3.4 Impact of ICT on metadata and metadata practitioners 84
  3.4.1 Consequences of ICT on metadata .......... 84
  3.4.2 Professional status of metadata practitioners .. 87
3.5 Current developments ..................................... 88
  3.5.1 Ontologies ........................................... 89
  3.5.2 Folksonomies ......................................... 93
  3.5.3 Use-neutral metadata and mash-ups ............ 96
3.6 Conclusions ............................................... 98

Summary: This chapter provides an overview of the technological evolutions which have influenced the creation and management of metadata within the cultural sector. The origins and the evolution of the term metadata itself is also debated, and the impact of ICT on their content and the status of metadata practitioners. Finally, three emergent approaches that each hold the promise to transform the creation and management of metadata are presented and evaluated.
3.1 Introduction

As Svenonius notes in the introduction to “The intellectual foundations of information organization”, there are two important reasons to study the history of the information sciences [137]. Firstly, most of the current problems have been around for the last century and a half or even longer and therefore it is interesting to see what solutions have been proposed in the past. Secondly, each vision of information indexing and retrieval expresses a specific ideology or world vision, that reflects the ambition of its creators.

This chapter starts with a critical historical overview of how ICT has been implemented within the cultural heritage sector, starting from the 1970s. From the early beginning, the sector had a number of expectations from new technologies, which sometimes still sound very familiar. I will investigate whether the expectations have been fulfilled by focusing on three central advantages that technology promised to deliver.

After confronting the past with current practices, the concrete impact of ICT upon metadata creation and management is analyzed. The growing importance of technologies has both affected the outcomes of documentation practices and the profile of practitioners. An overview of current developments closes up the chapter.

3.2 Mismatched expectations: a short history of the automatisation of cultural heritage documentation

The history of ICT applications within the heritage sector has already been treated in some publications [80, 135], but the focus lies exclusively on experiences in America in the 1960s, 1970s and 1980s. Therefore, I will mainly focus on the European situation and the developments from the 1990’s and early 2000’s.

3.2.1 Overview of the evolutions

Table offers an overview of the different periods and their evolutions. It is inspired by the tabel of John Perkins, but extended and applied to the European, and more specifically the Belgian, situation [116].

The automatization of catalogs started in some of the largest American museums and libraries at the end of the 1960s, by employing mainframe computers which were used by the industry for bulk data processing. A
3.2. MISMATCHED EXPECTATIONS

<table>
<thead>
<tr>
<th>Type of institutions</th>
<th>late 1960s</th>
<th>1970s</th>
<th>1980s</th>
<th>1990 - 1997</th>
<th>1998 - 2003</th>
<th>2004 -</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Largest institutions, mainly libraries</td>
<td>Largest institutions, mainly libraries</td>
<td>Large institutions with some museums</td>
<td>Museums and archives</td>
<td>On all levels</td>
<td>Cross-institutional collaboration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Large main-frames</th>
<th>Mini computers</th>
<th>Micro computers</th>
<th>Desktop computers</th>
<th>Maturation of imaging equipment and broadband Internet</th>
<th>Terrabyte storage and wifi</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Software</th>
<th>Custom-developed, low-level programming interface</th>
<th>Consortium development</th>
<th>First specific software</th>
<th>Large scale implementation of specific software</th>
<th>Web-based applications</th>
<th>Maturation open source software</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Functions</th>
<th>Sorting and searching records</th>
<th>Cataloging</th>
<th>Cataloging</th>
<th>Cataloging</th>
<th>Presentation</th>
<th>User interaction</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Issues</th>
<th>Hardware too expensive</th>
<th>Failure of a single solution</th>
<th>Standards</th>
<th>Standards</th>
<th>Digital preservation</th>
<th>Re-use of digital resources</th>
</tr>
</thead>
</table>

Table 3.1: Historical overview of the implementation of ICT within the cultural heritage sector

1968 survey from the International Council of Museums counted over 50 computer projects in the museology field in 13 different countries [48].

The equipment was so expensive that the technology was mainly used outside institutions. The project responsible for the automation of the thesaurus from the Institut Royal du Patrimoine de l’Art (IRPA) for example went to the offices of IBM to encode terms every week. Another option was to form a consortium to share the costs, like the consortium of New York museums that established the Computer Museum Network in 1967. Their goal was to initiate a knowledge center on computer applications within the museum domain. From the beginning, they also developed a collection management database to be used by all museums within the consortium. This database was designed for every museum to share its information with others, but was not very successful due to its generalist approach [135].
The goal of most of these automation projects was to search and sort through large numbers of records in a variety of fields. Mainframe systems had complex interfaces and the input and output of data were complex procedures [113]. Both mainframes and minicomputers used tape and disk units as storage mediums. The microcomputer, such as the Apple II or the TRS-80, finally made technology cheaper with prices ranging from 500 to 2,000 dollar. Secondly, they offered interfaces that could be manipulated by a broader range of people.

Apart from the hardware evolution, the most pivotal moment appeared with the rise of the Internet at the end of the 1990s. Again, the libraries were the first to embrace this new technology. Until 2000 most websites from cultural heritage institutions were just plain html pages which did not offer much more than static information regarding their collections. The library community, which has the longest experience with the use of databases for electronic catalogues within the heritage community, were the first to actually publish large amounts of metadata on the Internet through their Online Public Access (OPAC).

The possibility to publish metadata, and the digital surrogates of the resources, on the Internet gradually extended and transformed the functions of collection registration software. From the 1980s until the end of the 1990s, collection registration software was mainly used for auditing and inventory purposes. A limited number of commercial vendors such as The Museum System¹ or Adlib² propose since the 1990s out-of-the-box collection management software, although a majority of institutions used generic database management software such as dBase, Access or Filemaker Pro, to build their own custom-made applications. These applications mainly served the administrative and custodian needs of the heritage holders and researchers. The end user was only indirectly confronted with these databases, and had to rely on the staff of the institution to retrieve information.

However, the possibility to make the collection registration database searchable through the Internet obliged the database managers of cultural heritage institutions to also take into account the needs from a much broader range of users. This introduced a change within collection registration software, in the sense that it started to include more and more virtual exhibition features, which stand in a way opposite with the features of traditional collection management software. Table 3.2.1 as presented by Howard Besser [14] illustrates clearly the differences that used to exist between virtual exhibition and collection management software. The big disadvantage of the multimedia packages from the 1990s was their short life span, and thus relatively high production cost. By incorporating virtual exhibit-features within collection management software, collection

²See http://www.adlibsoft.com/.
3.2. MISMATCHED EXPECTATIONS

Managers try to incorporate tools within the database to create virtual exhibitions by reusing existing digitized resources and their metadata which are already available in the database.

<table>
<thead>
<tr>
<th>Interactive multimedia exhibition packages</th>
<th>Collection management systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designed for explanation and access</td>
<td>Designed for inventory control, record-keeping</td>
</tr>
<tr>
<td>Good user interface</td>
<td>Poor user interface</td>
</tr>
<tr>
<td>Narrative based</td>
<td>Object-based</td>
</tr>
<tr>
<td>Offer coherent view of some domain</td>
<td>No overall view of domain</td>
</tr>
<tr>
<td>Records are limited and carefully selected</td>
<td>Unlimited records, all are available</td>
</tr>
<tr>
<td>Frequently no database</td>
<td>Usually a powerful database</td>
</tr>
<tr>
<td>Limited user navigation</td>
<td>Unlimited user navigation</td>
</tr>
<tr>
<td>Always single user</td>
<td>Single or multi-user</td>
</tr>
<tr>
<td>Closed box system</td>
<td>Open or interoperable systems</td>
</tr>
<tr>
<td>Static</td>
<td>Dynamic</td>
</tr>
</tbody>
</table>

Table 3.2: Differences between collection and exhibition management software

The maturation of web applications also led to a breakthrough of web-based collection management systems. This implies that a local installation of the collection management software is no longer required as access to the database can be ensured by an Internet connection and a secure login. These web-based applications allowed to put the first step toward distributed collection management. Theoretically, the task of cataloging and indexing can now easily be outsourced to external experts if insufficient internal knowledge is available or if the sheer quantity of the images to describe is too large.

The database manager can administer read/write/update access at the level of an individual object and its individual descriptive fields. Large scale digitization projects such as the scanning of hundreds of thousands of historical postcards, newspaper clips or photographs can now recruit temporary employees, who receive a specific training to work directly in the database, independent of their location.\(^3\)

\(^3\)An exemplary project of this type of outsourcing was carried out for the Central Bureau of Genealogy of the Netherlands, when 50,000 scans of historical newspaper clips were described within 3 weeks by 20 temporary recruited persons who worked from their homes. Several metadata fields in relation to the person, activity/event and location depicted, the date of publication and the newspaper from which the clip was taken had to be transcribed from the scan of the newspaper article. To ensure optimal quality of the metadata, all these fields were encoded twice by different persons. The encoded values were then automatically compared within the database so that different versions of a certain metadata field were traced. In this way typographic and interpretative mistakes were avoided. This collection can be consulted at http://www.geheugenvannederland.nl/.
Web-based metadata creation also changes the temporal character of the encoding, in the sense that there is no longer a delay between the creation and publishing of metadata. This has not always been the case. A researcher in 1968 explains the metadata input procedure as following: “Approximately once every six months, the new items that have been keypunched are run on a card-to-tape routine and then merged into a master tape file. This master tape file is then sorted into four different hierarchical arrangements with a tape put from each sort. Each of these act as a computer printout, and the four listings together become the new index to be used until another is prepared” [33].

The latest important development is the maturation of open source collection management software. The restrictive software licensing policies of the commercial vendors and the lack of freedom to implement changes and updates into proprietary software promoted the development of open source metadata management tools [152]. Section 7.3.2 will present in detail an example of an open source collection management software.

3.2.2 Confronting expectations with reality

When reading the domain literature of the last 30 years, it is fascinating to confront expectations from the past with the current context and developments. Especially in the rapidly evolving field of information technology it can be interesting to take a step back and distinguish hypes from important novelties. The following three themes regroup the most common and widespread expectations of the last 40 years concerning what information technology could contribute for the cultural heritage sector.

3.2.2.1 Encompassing the information overload

Within the proceedings of the 1968 “Computers and their potential applications in museums” conference, most speakers noted in their introduction the inability of heritage institutions to manage the new flood of information. On the one hand, museums, archives and libraries were confronted with a lack of physical storage capabilities. But there were no technical possibilities to capture and store a digital version of the documents themselves, only their metadata. But hopes were expressed that this would become possible in the future.

More importantly, the growth of collections also already implied serious questions regarding the ability of heritage institutions to properly document all the new objects and documents. Much was expected from the computer to enhance the cataloging and indexing process, as the creation

---

4See for example [92].
and management of metadata within a paper-based world demanded a lot of manual copying and administrative work.

The hopes that heritage professionals had in digitization as a way of solving the problems of physical storage have been partly fulfilled. As the need to consult the original resources drops as more and more resources are digitally available, institutions are now increasingly storing original resources at other sites than the actual building of the institution. And in some cases, such as with historical newspapers, libraries even opt to destroy the original resources after digitization, which has led to an emotional debate.

The facilitation of indexing and cataloging by ICT however has not been as successful as heritage professionals had hoped in the 1960s and 1970s. The management of metadata and their exploitation for search and retrieval purposes has advanced at tremendous paces, but the actual creation of metadata is still a manual and cumbersome task. The automated indexation of images with CBIR techniques is unfortunately still limited to low-level semantics (see Section 1.3.1.3).

3.2.2 Attaining standardization and interoperability

The wish to standardize cultural heritage documentation with the use of ICT is one of the most recurrent themes throughout its short history. This long-standing focus on standardization and the resulting interoperability resembles in a sense the quest for the holy grail. In an article published in 1966 on the future development of a database for museums, the hope of interoperability was already expressed: “Fortunately few people want to set up a parochial closed system... Precipitious und unilateral activity in the museum world would squander a unique opportunity for international co-operation” [91]. There was a genuine hope that working with the same collection management software would automatically lead to standardization and interoperability. It is ironic to observe that exactly the same motives still prompt current policy makers to purchase collection registration software, even as local practices at the different individual institutions prevent standardization.

One of the first important events within the history of digital cultural heritage, the foundation of the Museum Computer Network in 1968, was built up around the idea of developing a uniform database which would allow different museums to work together and exchange information. Several networking groups and software user groups were created in the 1970's in order to establish national standards for metadata sharing [135]. The project unfortunately never managed to deliver results. In general, libraries are the only institutions that have actually managed to reach a functioning level of interoperability. The two reasons behind the success
within the library sector is the non-unique character of their resources and the financial incentives for copy-cataloging (see Section 4.3.1).

Progress has been made recently by adopting a more pragmatic approach toward interoperability. Up until the beginning of this century, a considerable number of projects tried to develop a central database to which all participants had to conform their metadata schema. This approach has showed to be too cumbersome and has been replaced with a more decentralized vision of metadata sharing. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a perfect example of this new approach. The Dublin Core metadata scheme is used to map different metadata elements to a common structure. This merging of information often results in poor metadata quality [55, 54].

3.2.2.3 Promoting interactivity

Along with interoperability, the notion of interactivity is one of the key advantages that is attributed to the digitization of cultural heritage. The digital medium effectively offers unprecedented possibilities for the interaction between a cultural resource and a user. The fact that a user can zoom in, pan and crop images was considered a major characteristic of the digital age [15]. But as Simon Knell notes, this much hailed interactivity often turns out to be a process which does not lead to a deeper understanding:

Interactivity was a museum buzzword in the 1970s and 1980s, but it too has been questioned, particularly with the rise of modern edutainment centers where there is often no logical point to the interaction and no relationship between action and outcome, and where - because there are no real objects - intellectual depth is illusive [84, p. 141].

The Minerva working group, which is funded by the European Commission, has recently published a report on good practices regarding user interaction in the context of cultural heritage digitization projects [29]. The report offers an extensive overview of the new possibilities offered by web2.0 tools, but does not offer a critical review of the added-value of these services. Section 6.4 will elaborate on this issue in the context of user comments, and identify where exactly lies the value of the interaction between a cultural resource and its users.

3.3 Metadata: old wine in new bags?

When studying the development and evolution of metadata as a research topic and a professional practice, an obvious question that rises is where
3.3. METADATA: OLD WINE IN NEW BAGS?

to position the traditional practices of cataloging and indexing within the metadata discourse. Is it just old wine in new bottles, meaning that metadata is just a trendier word to give a sexier image to cataloging and indexing?

A victim of its own success, the term metadata is currently being used within a large array of application domains. These each give their own interpretation to what exactly metadata are, which results in different interpretations, uses and levels of expectation. Librarians identify metadata with their highly structured and professionally created catalog, where as webdesigners consider metadata as useless keywords to be entered at the beginning of a html file, as the keywords that are included within the meta-name headers no longer are indexed by search engines due to spamming practices. Even domains that have no direct interest in information retrieval have incorporated the term within their vocabulary. The contemporary art scene for example has embraced metadata as an inspiration source and point of departure for exhibitions and art works.

3.3.1 Overview of the evolution of metadata

Jane Greenberg published in 2005 one of the only existing overviews of the historical context of metadata, which offers a comprehensive summary of the different developments and definitions of metadata and metadata schemes [67].

Initial context of use

The term itself was coined by Jack Myers in 1969, and registered in 1986 as a trademark for his enterprise, The Metadata Company. Boydens situates the first metadata related research within the domain of statistical databases, from where the use and application of metadata spread out to other domains in the 1980's [25]. Within this early context, metadata were understood and used in the sense of meta-information systems, which have the intention to reduce the level of uncertainty in databases by providing contextual information on the content of databases.

5Spamindexing tries to influence the ranking results of webpages by search engines, by including popular keywords in the meta-headers of a html file. As a result of these practices, search engines no longer use the content of the meta-headers but look for other criteria such as the link-structure of webpages to rank search results. Spammers have then reacted by including false links on their pages. See [40] for an overview of spamming and the measures that search engines take to respond to these practices.

6See for example the article of Bart De Baere, that incessantly refers to metadata, or the exposition Information/Transformation, that was held in Antwerp in 2005, where the groundbreaking work of Paul Otlet, one of the grounding fathers of information science, stood central (http://www.extracity.org/info_transfo/index.htm). The cupboards with the catalog cards from his institute also appeared in the exposition "Visionair Belgie", curated by Harald Szeemann in 2005.

Our understanding of metadata focuses more on the attributes of individual resources, that can, or not, be contained within a database. The rise of the Internet since the second half of the 1990’s as the major tool for resource discovery catalyzed the need for standardized and machine-readable metadata throughout different domains. This gave rise to the development of the first metadata standards, such as the Federal Geographic Data Committee (FGCD), which states how resources should be described with which type of descriptive fields.

During the same period, the Dublin Core Metadata Element Set (DCMES) was developed in 1995 by the OCLC. Its mission is to support resource discovery on the Internet and has seen a wide uptake as a standard. The international research community that supports its development concentrates on finding the best trade off possible between interoperability, ease of use, and semantic richness.

**Adoption within the business administration**

The Gartner report on the hype cycle for data management gives an overview of the adoption of new technologies and standards within the business administration domain. The hype cycle for data management provides an insight into which technologies are already mature and deployed on a large scale within the industry. Figure 3.1 represents the five stages that every new technology that is being adopted follows. After the initial technology trigger follows a high peak of expectations, which is then logically followed by a period of disillusionment. This phase is then followed by a longer period where the technology is gradually proving its actual use, from where on it moves to a more stable plateau of production.

Figure 3.1 shows us that metadata are at the very starting point of the hype cycle. This illustrates that within the business administration domain, metadata management has been neglected in favor of the adoption of more purely technological tools. The use and management of metadata is thus at a very immature level of implementation within most large scale enterprises.

However, the Gartner report points to a high benefit rating for metadata management and recognizes the strategic importance of metadata for the

---

8 Largely pre-dating the Internet, commercial database servers such as DIALOG already published very large quantities of data via a distributed platform. These kind of data that are enclosed within databases are now known as the so-called “deep web” [26].

success of many IT-related projects. With the help of Figure 3.1, Gartner points out that they will augment their attention toward metadata and predict an increasing importance of metadata within the business administration domain [142]. Nevertheless, it is important to note that the implementation of metadata strategies within large scale enterprises will be a process that takes many years.

Figure 3.1: Hype cycle for metadata management as published by Gartner

The development of metadata schemes

The work on Dublin Core has led to an increasing awareness of metadata throughout different application domains, from which point an uncontrollable number of domain-specific metadata element sets were published. These are commonly referred to as metadata schemes. As Greenberg notes, there exists no commonly accepted definition of what exactly a metadata scheme is, but I will refer here to the definition of Priscilla Caplan as a “set of metadata elements and rules for their uses that has been defined for a particular purpose” [31]. The rules that accompany the metadata scheme specify how each element should be interpreted and

---

10 Confusion can occur with the use of scheme in the sense of classificatory and terminological systems, such as vocabulary and syntax encoding schemes. Joseph Tennis refers to my understanding of a scheme by using the term schema [140].
used, and can also contain links to controlled vocabularies or specific data formats. In relation to the database world, one could compare a metadata scheme with a database model, and the metadata specifications with the integrity constraints.\textsuperscript{11}

There are almost as many typologies of metadata schemes and elements as there are metadata schemes.\textsuperscript{12} The following four types of metadata are commonly identified throughout different typologies:

- **Discovery metadata**: allow the identification and the retrieval of a document, such as the creator, the title or the subject
- **Use metadata**: facilitate the technical and intellectual use of the document. The technical use encompasses for example the type of format, the intellectual use includes the publication rights.
- **Authentication metadata**: make the evaluation of the authentication possible, such as information regarding the version of the document
- **Administration metadata**: help to structure information regarding the management and conservation of a document, such as the date and method of acquisition. These can also incorporate meta metadata, information regarding the metadata, such as the name of the indexer.

These typologies are important to understand the different functionalities supported by metadata, but in practice the different types blend into one another. Ingrid Mason for example distinguishes curatorial (keeping practices), semantic (terminology and documentation) and technical (web design and communication) metadata [101]. But the EAD, which she files under technical metadata, also incorporate essential semantic information on the resource, where as the DCMES, which is classified as a semantic standard, can also contain technical metadata. No single metadata typology has managed to draw clear and stable lines between its different types.

As Greenberg mentions, the development of a metadata scheme is very much an empirical activity, in the sense that experts from a specific domain gather to reach a consensus on the elements and specifications to be used for a scheme, based upon their professional experience that ideally is a mix of technical and content skills.

What exactly is the impetus behind the development of this wide variety of schemes? Ingrid Mason notes that the elaboration of metadata standards are international, collaborative ventures, driven by shared interests of providing support for the local and global management of digital resources [101]. The development of a metadata standard is a very resource intensive activity, which is commonly financed within the framework of a

\textsuperscript{11}See the standard work of Elmasri [49] on database design.

\textsuperscript{12}See for example [31], [89], [62], [66] and [101].
national or international project, drawing its funding from organizations such as the Institute of Museum and Library Services (IMLS) for American projects, or from a program of the DG Information Society of the European Commission.

Unfortunately, this project-based funding implies some danger for the sustainability of the newly developed standard. If the work on the endorsement and implementation of the standard is not overtaken by one of the project-partners, then chances are quite high that the standard goes into oblivion.\(^{13}\) It is hard to keep track of all of the metadata schemes that have been developed or that are currently in development. Some initiatives, such as the SCHEMA-project, provide a repository for metadata schemes and their successive versions.\(^{14}\)

### 3.3.2 Specific metadata schemes for the cultural heritage domain

The complexity and the different intertwining levels of description of metadata standards can also be illustrated by taking a closer look at the different standards and schemes that were designed in the last decade specifically for the cultural heritage community. Examples of the different levels of metadata standards and schemes illustrate the “russian puppets” syndrome I already mentioned when addressing the ever-extendability of metadata:

- **Local metadata schemes**: With local metadata schemes I mean schemes that are specific and rich enough to grasp all of the information that needs to be recorded in order to fulfill the local needs of a cultural heritage institution. Here for example, I can refer to the SEPIADES scheme, which was specifically developed for digitized historical photographs. With its 400 elements, it proposes a very rich model, which also makes it difficult and costly to apply.

- **Global metadata schemes**: Global metadata schemes take on the role of providing crosswalks between different local metadata schemes. The Dublin Core is the most well-known example. In theory, the minimal set of 15 Dublin Core fields could be used to natively describe cultural heritage resources, but in practice they do not offer enough possibilities to fully describe a resource. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), uses for example

\(^{13}\)The SEPIADES project is unfortunately a good example of how a metadata scheme can die a slow death. This European project developed a metadata scheme, a software tool and guidelines to document photographic collections, either in their analogue or digital form. The metadata scheme is very rich, offers multi-level description and over 400 elements. Despite the quality of the project, the standard and the software tool practically not being used by the institutions the scheme was specifically developed for.

\(^{14}\)See the website [http://www.schemas-forum.org/registry/](http://www.schemas-forum.org/registry/).
Dublin Core fields to map the fields from different content providers that have more or less the same semantic meaning to one single field.

- **Container metadata standards:** Different metadata schemes are sometimes used for the management of metadata of one single record. So called container metadata standards offer a framework to simultaneously use different schemes, that each have their specific use. The Metadata Encoding and Transfer Syntax (METS) is currently the most-used container metadata scheme, and is a “standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library”. Concretely, this means that the same bibliographic record can both incorporate ONIX (the metadata standard for book industry product information), which will be used by a publisher, and a Machine Readable Cataloging (MARC) package that will address the needs of a library. Figure 3.2 shows a METS record where one container holds a Metadata Object Description Schema (MODS) record, consisting of a translated MARC record from the University of California union catalog, while another holds a record called “ucpress”, consisting of bibliographic metadata from an in-house database at the University of California Press.

```xml
<?xml version="1.0" ?>
  + <mdHdr>
  + <dmdSec ID="mods">
  + <dmdSec ID="ucpress">
  + <amdSec>
  + <fileSec>
  + <structMap>
</metas>
```

Figure 3.2: Collapsed view of a METS record, illustrating the concept of container metadata standards

- **Conceptual metadata schemes:** These schemes do not deliver a limited set of fields that can be immediately implemented within a collection registration software, but provide rigorous definitions and a formal structure when working on the development of a metadata

---

15 http://www.loc.gov/standards/mets/
16 http://www.editeur.org/onix.html
17 http://www.loc.gov/marc/
18 http://www.loc.gov/standards/mods/
19 Example cited from [139].
scheme and strategy. The Conceptual Reference Model from the International Committee for Museum Documentation (CIDOC-CRM) does exactly this for the cultural heritage sector, by providing definitions of concepts and relationships between concepts. Figure 3.3 illustrates how the different concepts that intervene during the acquisition of an object, such as an actor who can be the current or former owner of an object. This scheme, which also indicates the cardinalities of the relationships in between concepts, can then be used to design a collection registration database.

Figure 3.3: Modeling acquisition information based upon CIDOC CRM

Extensive overviews of all of the metadata schemes and standards that have been developed for the cultural heritage world already exist, so therefore I do not provide a thorough overview here. Apart from the fact that overviews already exist, they also tend to be outdated quickly. It is exemplifying that a significant part of the external links toward metadata standards that are included within the website of the Canadian Heritage Information Network (CHIN) are already broken.

---

20 The modeling schema was copied from the CIDOC CRM wiki page on http://139.91.183.17:81/tiki/tiki-index.php.
21 One of the most recent and thorough articles is [101]. The website of the Canadian Heritage Information Network (CHIN) also provides an extensive overview of cultural heritage standards (http://www.chin.gc.ca/English/Standards/index.html).
22 See the links to the “Méthode d’inventaire informatique des objets beaux-arts et arts décoratifs” and the “RLG REACH Element Set.”
CHAPTER 3. APPLYING ICT FOR METADATA CREATION: AN OVERVIEW

3.4 Impact of ICT on metadata and metadata practitioners

This chapter focuses on a review of the different technologies used for the documentation of cultural heritage. As the framework of the stratified time demonstrates (see Section 1.2), there is an interaction between the intermediate time, to which the technological evolutions discussed in this chapter correspond, and the long and the short time. Therefore, I will analyze here how the technological evolutions as described in Section 3.2 affect the metadata themselves, which correspond to the short time, and the evolution of the profile of metadata practitioners, which is analyzed in the long term.

3.4.1 Consequences of ICT on metadata

Before I focus on the impact of ICT, I can also refer to evolutions from the paper-based era of documentation practices. Bowker and Star note that even the type of paper influences the content of classifications. The original International Classification of Diseases (ICD) for example consisted of two hundred disease categories as this was the number of rows in the standard Prussian census form [24].

Other pre-computer transitions have also been illustrated within the work of JoAnne Yates on late 19th century office technology. Correspondence books used to collect in a chronological order copies of all outgoing letters from an enterprise. Technological evolutions that now seem trivial, such as the invention of the manila folder and the hanging file drawer, offered the first possibilities to escape the strictly linear way of organizing information and documents in the book form, by introducing the notion of subject access. The later invention of carbon paper to easily create multiple copies of the same document, allowed the storage of the same document in multiple ways and places [23]. The card-based catalog offered unforeseen ways of offering multiple access to the same reference, by the possibility to create several cards for the same bibliographic reference. However, there were physical and economical limits to the number of cards produced for one reference: the so-called “rule of three” limited the number of cards based upon subject headings and the author for one title to three [75].

These limits on the extent of cross-indexing disappeared with the introduction of computers. However, limitations on the number of metadata fields and their content were not resolved, as the management of memory remained until the 1990s one of the key factors to be taken into account when developing large databases. These limits can have unforeseen consequences on the quality of the encoded metadata. Due to the limitation of the number of characters that could be entered in a certain database field,
the former librarian of the Royal Museum of Central Africa entered the rest of the metadata in other fields, which were not destined for this type of information. This resulted in the corruption of 10,000 bibliographic records [150].

**Collection registration software as black boxes for metadata creation**

The historical overview of the ICT developments within the cultural heritage sector marked the widespread uptake of collection registration specifically developed toward heritage institutions in the middle of the 1990s. As such, this was a positive development as it catalyzed the creation of metadata. However, these proprietary software packages did not always have a positive influence on the quality of the produced metadata, as the metadata scheme and the constraints often did not match the specific needs of an institution.

This led to situations where metadata practitioners adapted their descriptive needs to the possibilities offered by the software. As a librarian points out: “We, as resource describers, have not demanded alterations to the software. Instead, we have altered our practices to fit in with the software” [133]. This problem area is unfortunately largely unacknowledged.

Most of the collection registration software that is currently in use within the cultural heritage sector does offer ways of customizing metadata schemes or of specifying integrity constraints. However, the customization of generic collection registration software in practice tends to be a very problematic. Software such as The Museum System (TMS) [23], which is heavily used by large museums throughout Europe and Northern America, is more or less a black box which can only be modified through the software developer itself or a small number of technical consultants, who charge heavily for any customization of the database and the interface.

Other software such as Adlib Information Systems (Adlib) [24], which is broadly used in Belgium and the Netherlands, does allow the customization of the metadata fields, but only a small minority of libraries or museums has the in-house competencies for this kind of customization. This results in situations where an institution is confronted with an interface that holds 300 to 400 metadata fields, from which it uses only 20 or 30. And since the institution does not know how to add or customize the default metadata fields, some very specific metadata of the institution are contained within a default metadata field that does not correspond with its content. Figure 3.4 illustrates the interface of the Adlib museum module as it is used in the Archief en Museum voor de Socialistische Arbeidersbeweging (AMSAB) [25]. As one can see on the screenshot of the collection

---

23 See http://www.galleriesystems.com/products/tms.html
24 See http://www.adlibsoft.com/
25 See http://www.amsab.be
registration database, only a minority of the metadata fields are being used.

Figure 3.4: Screenshot of the collection registration database of AMSAB, with a majority of empty metadata fields

An significant part of the problems that will be identified in chapter 5 result from the lack of adequacy of generic collection registration software toward the local descriptive practices. I can refer here to that chapter for more concrete examples.

The “black box” character of proprietary software and the resulting difficulties to implement modifications have led to the development of open source alternatives [152]. Especially the library sector now has two mature and well-used open source tools, which are Koha[26] and Greenstone.[27] The museum sector used to lag behind, but now has with CollectiveAccess[28] a full-featured collection management system. Section 7.3.2 will present this software more in detail, as one of three strategies for enhancing metadata quality from the dissertation is implemented within this open source software.

---

[26] See http://www.koha.org
[27] See http://www.greystone.org
[28] See http://www.collectiveaccess.org
3.4. IMPACT OF ICT ON METADATA AND METADATA PRACTITIONERS

3.4.2 Professional status of metadata practitioners

The indexing and cataloging of art were in the 19th century heavily influenced by the notion of “ekphrasis” and thus considered an art or a science. With the professionalization of the heritage institutions in the first half of the 20th century, the professional identity of librarians, curators and archivists slowly became more established.

Completely the opposite movement took place the last decade within the cultural heritage sector. The rising popularity of the notion of metadata in the last ten years has been accompanied with a growing uncertainty on which competencies metadata specialists should exactly have. Despite the popularity of “metadata” in job titles, there is very little consistency in the skills this profession exactly requires [56]. This uncertainty and ambivalence is also reflected in the redefinition of the scopes and the curricula of Information and Library Science schools. Section 3.2.4 will go into further detail regarding the reforms of digital libraries curricula.

An important effect of the automatization of the documentation practices is the shift in the profile of the professionals who perform metadata creation and management. Within the museum sector for example, curators traditionally have been responsible for providing documentation. The appraisal and description of archives has also been performed by higher level personnel.

The introduction of collection registration software in the Institut Royal du Patrimoine de l’Art (IRPA) led to a situation where some of the art historians refused to enter metadata into the computer. They considered the encoding of information into a database as a purely administrative task. The older generation of the scientific staff also expressed a feeling of discomfort that their scholarly or curatorial knowledge of the collection, by entering it into a collection registration database, was somehow thrown to the masses.

The challenges implicated with the growing number of documents to be described have led to a situation where it becomes very difficult to find a balance between the level of description needed and the available time to be invested. This tension was perceived as early as in 1968:

We might trim our sails down and predicate our system on the lowest level of cataloger-help, working on the simplest kind of information. The sacrifice of leveling down may be critical, however, because the precision and quantity of output information is directly proportional to the complexity of input informa-

29The phenomenon of de-professionalisation is concentrated in non-strategic sectors, such as my application domain of the cultural heritage field. In the domain of the social security administration for example, it is essential to have highly trained documentation specialists that create and maintain the documentation of the administrative databases.
tion. At what point will cutting down complexity and precision jeopardize the value of the computer to us? [92, p. 45]

An important factor within this discussion of the professional status of metadata practitioners is the financing of digitization projects. As discussed in Section 2.5, the metadata creation and management is performed by professionals who have a short-term working contract. The frequent change in personnel definitively affects the consistency of metadata quality. Especially when working with images, the lack of inter-indexing consistency can become an important issue [81]. And here we see how changes in the long term in the form of the evolution of the financing policies, can have an indirect impact upon the metadata quality.

3.5 Current developments

Ontologies and folksonomies have both been hailed within their respective research communities as revolutionary new types of metadata. However, both terms are very vague and little consensus exists on their actual definition and scope. The work on ontologies mainly takes place within the academic semantic web community (see Section 1.3.1.3). Folksonomies do not really have a scientific basis, but were launched and promoted by the blogging community. The publications that have been dedicated to the topic mainly come from the new media studies and social sciences.

The interest in analyzing ontologies and folksonomies together lies in the fact that they both demonstrate the new and extreme limits of the indexing languages landscape and illustrate distinct epistemological traditions within the information science field:

There has always been evident within the study and management of information a schism between two different world views. On the one hand is the classical view of logic and digital computing which focuses on the underlying order of information. On the other hand is the view of truths as culture-bound, knowledge as culturally dependent and information as a manifestation of power [145, p. 156].

Ontologies represent the influence of formal logic and computer science, which try, by concentrating on the structure of information, to formalize and categorize reality. Folksonomies on the other hand stand for the influence of the humanities, such as history, semiology and cultural studies that point to the cultural aspect of information and the importance of the context within which a document is created and interpreted.
### 3.5.1 Ontologies

The term “ontology” has been borrowed by computer scientists from philosophy, where an ontology is considered a system of categories that represent reality. The research within computer sciences on ontologies needs to be situated within the larger context of the semantic web. By representing metadata in a strongly formalized and standardized manner, the semantic web community strives toward a generic interoperability between resources. But as Fabien Gandon noted in “Le web s’mantique n’est pas antischolaire”, very little consensus exists on the exact scope of the semantic web [57]. The official W3C mailing list concerning the semantic web is for example frequently monopolized by endless discussions regarding the ambitions of the research area.

The current implementation of the semantic web is mainly based on two standards: the Resource Description Framework (RDF) and the Ontology Web Language (OWL). As its name implies, RDF offers a framework for the description of a resource, in the format “subject - predicate - object”. This datamodel allows the formulation of simple and disambiguated declarations, that can be combined and layered with other RDF expressions. URL's offer a unique identification of each part of an RDF expression. As a predicate between a book and an author, the Dublin Core field “creator” can be used, which can be identified with the URL http://pur1.org/dc/elements/1.1/.

However, the RDF syntax is quite limited and does not allow rich expressions. OWL, which delivers the actual building blocks of an ontology, allows the definition of supplementary constraints that are too expressive for RDF. OWL consists of three components: first of all, classes represent individuals that share the same characteristics. Secondly, properties define facts or relationships between classes. Finally, instances are individuals belonging to a class that take up characteristics defined by properties.

What are the advantages of ontologies? Firstly, they allow more expressivity in the metadata scheme. The ABC ontology for example formalizes temporality in a way that allows far more expressivity [88]. Metadata schemes are normally constructed around a resource which is considered to remain stable throughout time. This assumption becomes quite problematic in the case of the digitization of historical photographs. How does one describe for example a glass plate negative from 1890, from which a print was made in the 1940's and that was scanned in 2003? Secondly, ontologies allow automatic reasoners to test the consistency of the ontology in question and to infer the hierarchy between classes. And finally, by formalizing the relationships and constraints between the concepts of a certain domain, an ontology can guide the process to develop metadata.

---

30 See http://www.w3.org/RDF/
31 See http://www.w3.org/2007/OWL/wiki/OWL_Working_Group
and database schemes. For the cultural heritage sector, the CIDOC Conceptual Reference Model offers definitions within a formal structure for the description of concepts and relations that are used within the documentation of cultural heritage. By offering this ontology, CIDOC hopes to promote interoperability between different collections. Section 3.3.2 and Figure 3.3 give more details on CIDOC.

Despite these advantages and the large scientific interest in the semantic web, it still hard to put one’s finger on the exact added-value of ontologies in practice. The ultimate goal of these operations is, like one of the developers behind the E-Culture project describes, to “have a multitude of collections that will become a part of the interoperable space” [143]. This primal focus on interoperability becomes an end in itself, and tends to make abstraction of the local descriptive needs of an institution.

Quite some prototypes have been presented the last five years to implement Semantic Web techniques within image metadata management tools [70]. However, most of these research projects have a short life span and are never implemented in the day-to-day practice. The creation and maintenance of ontologies demands a lot of resources. Everyone who has hands-on experience with the development of a thesaurus knows how time-consuming and intellectually challenging the creation of a controlled vocabulary is. Only large institutions that have the financial means to invest in such a long term project have a chance to successfully develop and update a thesaurus.

Considering the higher complexity of ontologies and the limited resources of the cultural sector, it seems highly unlikely that we will be witnessing a widespread adoption of ontologies in the cultural heritage sector. The return on investment is simply too small to defend the costs toward financial backers. This is also the case with relational databases. This technology has been around for some 30-40 years, but apart from the library sector, very few cultural institutions have fully taken advantage of relational databases. The problems noted in Section 3.4.1 illustrate the difficulties the cultural sector has to implement this consolidated and mature technology.

The utility for the end-user is also questionable. Complex queries can be conducted with the help of ontologies, but if one considers that most end-users already have problems with boolean queries, then it is rather doubtful if they could take advantage of the added-value of ontologies. A pragmatic solution would be to hide the complexity of ontologies and to display only a simplified version in the user interface. Ontologies are also often criticized for their ambition to formalize and objectify, and, by this operation, to reduce the multi-faceted reality. The specific context of the production and reception of a resources is thus also not acknowledged.

Finally, I would like to relativize the sometimes naive and hype-driven discourse on the application of semantic web tools, while devalorizing the
3.5. CURRENT DEVELOPMENTS

indexing and cataloging work that has been done in previous decades. Cameron and Robinson in their chapter on the postmodern approach toward museums documentation criticize traditional documentation practices as an “empiricist tradition”, where experts assign “disciplinary classifications, descriptions, and anonymous statements of significance and provenance” to collections [30].

However, a few pages further in their chapter, they put their hopes in the semantic web, where “self-describing” resources will be endlessly interconnected with other resources. What Cameron and Robinson apparently do not realize, is that the formal descriptive practices they rail against, form the necessary foundation of metadata needed to move toward a better semantic web. A good example is the aquabrowser interface of the catalog of the public Flemish libraries, illustrated in Figures 3.5 to 3.7, which offers a very semantic web like interface. This trendy semantic-web flavored result is only possible because of 20 years of shared quality cataloging.

Regardless of the critique on the hyped nature of the semantic web, some concrete projects have really resulted in the augmentation of metadata quality, by linking institution specific metadata to standardized metadata schemes and controlled vocabularies. Especially the AnnoCultor tool\[32\]

\[32\]See http://annocultor.sourceforge.net/
and its implementation within the E-Culture Multimedian project\(^3\) offers one of the most interesting and concrete semantic web projects on cultural heritage metadata. Within this project, metadata and controlled vocabularies from several large cultural institutions are semantically interlinked together, by mapping the specific schema of the metadata of each institution to standardized RDF schemas based upon Visual Resources Association (VRA) Core, which is a specialization of Dublin Core for images. All of the controlled vocabularies, consisting of custom-made geographical or person names, are converted to the Simple Knowledge Organization System (SKOS) and then aligned to some of the most important and well-used controlled vocabularies such as the Art and Architecture Thesaurus (AAT), the Union List of Artist Names (ULAN) and the Thesaurus for Geographic Names (TGN).

This mapping cannot be performed automatically, but a collection of rules was written to map the metadata schemas and the controlled vocabularies. From a data modeling point of view, it’s interesting to see how the RDF conversion of an XML description breaks the record apart in a triple, being an instance of the VRA Core class vra:Work and an instance of the VRA Core class vra:Image, which are linked together by the property vra:relation.depicts.

The researchers provided some figures on the success rate of the mapping between the local and the global controlled vocabularies. The success rate differs largely from one list to another, which depends on the local uniqueness of the collections presumably.

The developers estimate that the conversion of a dataset from a middle

\(^3\)See http://e-culture.multimedian.nl/
to large-sized museum may take up to one work month for a skilled programmer that has a general knowledge of the collections he is treating.

### 3.5.2 Folksonomies

The neologism “folksonomy” is used to designate the outcome of the description of online resources by its community of users. This description happens in a completely uncontrolled manner, by according freely chosen keywords to websites, music, videos and images through a simple interface. The totality of keywords used by the user community can then, in a second step, be used as a search interface, for example in the form of a tag cloud, which groups the most popular tags (see Figure 3.11 for an illustration of a tagcloud). This way of document indexing is mainly applied within the blogging community or on social network websites. It greatly differs, by its libertarian character, from traditional indexing methods, and has stirred a debate between partisans and opponents of controlled vocabularies.

Some cultural institutions, mainly in America and Australia, are experimenting with social tagging on their websites and collection management systems. The technology is quite easy to put into practices and requires no special intervention by the staff. In the museum sector, the most well-known project is the Steve-project, in which institutions such as the Metropolitan Museum of Art and the Guggenheim participate. See Figure 3.10 for an illustration of the tagging interface within a collection registration database.

The Archives of Normandie present on the image sharing site Flickr a part of their digitized collection of World War II photographs. Users are invited to tag and comment on this collection, as the archives themselves do not have sufficient resources to ensure the description of the collection. Several prototypes have also been developed in the library world, within the OPAC for example of the Penn State University State library. The open

---

34 See http://www.steve.museum/
source library cataloguing software PMB has also incorporated a social tagging tool within its software, which allows users to tag references.

What is the long term value of social tagging for the cultural heritage community? The most common criticism regarding folksonomies points at the non-controlled character, which results in problems related to synonymy and polysemy. But some more fundamental comments can be made, especially if we look at the different emmanations of user-generated metadata of the last years.

The idea of giving the user the possibility to facilitate the access to a certain resource is not as revolutionary as some folksonomy advocates presume. Firstly, the author of a webpage can enter in the html-document keywords using the specific meta html-tags. Secondly, the pagerank mechanism of Google that influences the order of search results is partly based on the presence of links to a certain webpage, found on other webpages. By publishing a link on a website, users already can actively influence the access to a webpage. But the relevance of both systems, the keywords and
3.5. CURRENT DEVELOPMENTS

Figure 3.11: Screenshot of the tagcloud of the Archives de Normandie project, representing the most popular tags.

the weblinks has been lost by the spam-indexing phenomenon. And it is thus probable that, when folksonomies become more popular, the same problem with spamming will appear within folksonomies.

Social tagging also completely changes the regular documentation cycle. Traditionally, the classification of information preceded the access to the same information. But as Ertscheid and Gallezot observe: “Jusqu’ici, l’accès était globalement subordonné au classement. Or c’est désormais le classement qui pourrait être subordonné à l’accès.” [51]. Tagclouds, which are built up with keywords or tags that are the most popular amongst users, and as such are being used to build up a dynamic interface.

Another remark is closely related to this last point. One of the first principles in cataloging and indexing is the wish for objectivity. In practice, this can never quite be attained, but it is still considered a guideline. Social tagging on the contrary explicitly incorporates keywords that are closely, or even solely, related to the indexer. Examples of such tags are “toread” or “beautiful”. Although this adds an interesting social component to in-
formation retrieval, this kind of egocentric indexing is only interesting in certain situations. This subject will be more discussed in chapter 6.

In the end, social tagging and folksonomies should perhaps more be interpreted as statements about how content holders wish to empower their users, and not so much as effective information retrieval tools with intrinsic value. In “Memory practices in the sciences” Geoffrey Bowker points out that an underestimated part of our memory practices simply allow us to manage and frame the present, without the actual intention of using the created archive to access the past. In this sense, memory practices are more a “way of acting on the present rather than recalling the past” and “the act of rendering memorable does not mean that at any stage it will be remembered” [23].

In parallel with this thinking, folksonomies and social tagging can be seen as a way of communicating between an institution and it’s users, within the actual climate where user-participation is highly regarded. The main function of tagclouds is more to be found in this symbolic function of how a museum or an archive wants to emancipate its userbase, than an actual search tool. Again, this is an example of how mentalities which belong to the long time influence the actual metadata creation process.

3.5.3 Use-neutral metadata and mash-ups

The previous sections and chapters have described how institutions are trying to adapt their descriptive practices as much as possible to the needs of their users. A recent development within this discussion is however to adopt and radicalize the idea that an institution can never predict user needs, and should therefore concentrate on offering data and metadata in a “use-neutral” manner. From a technical point of view, this approach toward metadata quality is based upon the development of Application Programming Interfaces (APIs). These applications allow external parties to access data and their metadata, so that they can develop new services upon existing data and metadata.

The idea behind this approach is that content providers allow and even encourage users to “hack” data and metadata by offering them a toolkit which allows them to re-use data and metadata that can be re-used to fulfill their specific needs. This approach has been labelled with the term “mashups”, referring to the practice of dj’s to create new music by mixing existing music from different sources.

A somehow trivial, but real-life example, of where the mashup idea can concretely lead to are the publications of Philip M. Parker. This researcher, who holds the Chair Professorship of Management Science at INSEAD, has published over 85,000 books, most of which can be order on Amazon.com. Examples of titles he has produced as an editor include for example: “The 2007 Import and Export Market for Electrical Relays Used with Circuits
3.5. CURRENT DEVELOPMENTS

of Up to 1,000 Volts in Ireland", “The 2007-2012 Outlook for Lemon-Flavored Bottled Water in Japan”, “The Official Patients Sourcebook on Hemochromatosis”, “Websters Persian (Farsi) to English Crossword Puzzles: Level 5”. These obscurely titled books contain between hundred and three hundred pages and are available from Amazon for prices ranging from 200 to 700 dollar [18].

In fact, Parker developed a method to assemble automatically diverse content from the web into the format of a book, which can be produced in a matter of two hours. Entire pages are copy/pasted from existing sources that are free from copyright, but references are included to the original sources. Digital print on demand makes it possible to only produce a book when it is actually purchased, and at very low costs.

Several examples can be given of this growing re-use of metadata within the specific context of the cultural heritage domain. The Rijksmuseum in Amsterdam has launched in 2005 the Rijksmuseum widget. A widget is a simple application a user can install on his computer to perform a a very specific and limited task. When launching the Rijksmuseum widget, the user is presented a random artwork from the collection. Clicking on the image flips it to a display of the metadata of the image. See Figure 3.3 for an illustration of both sides. A direct link is provided to consult the online record that describes the object within the object registration database. Shortly after the launch of this widget, an student hacked the non-public XML stream to offer an Rich Site Summary (RSS) feed that is now know as the informal museumfeed of the Rijksmuseum. This example illustrates the opportunities developers have to take existing data and to model them in a different way.

Another example is offered on the website www.chopac.org, which provides a tool to export the descriptions of books from the Amazon database to the MARC format. Google Maps probably offers one of the most well-known and -used APIs. The service is increasingly used to add a geographical component to the search- and browse interface of cultural heritage collections. Figure 3.12 illustrates how Google Maps is integrated within the collection registration software of the Coney Island history project, which makes it possible to publish a map on the website that provides links to historical photographs from the database. This feature allows users to intuitive browse through collections and to compare the actual situation of a site with historical images which display the former state of the site.

35 See http://breyten.livejournal.com/111482.html
36 See http://www.coneyislandhistory.org/development/
3.6 Conclusions

The historical overview of the first part of this chapter identified the most crucial technological evolutions of the last decades for my application domain. The confrontation of the three key hopes that were formulated at the dawn of the digitization of metadata in the 1970s with the current situation obliges us to relativize the positive impact of ICT. Concrete examples have showed that two major problem areas from the early stages of digitization, being the interoperability of metadata and the management of the information overload, are still far from resolved. However, the development of the Internet has resulted in a tremendous boost of the interactivity between cultural heritage resources and their users.

The second part of the chapter investigated the concrete results of the technological advances on metadata quality. Here the focus was laid on the largely unacknowledged “black box” character of collection registration software, which results in the inadequacy of the metadata schema to the specific needs of the cultural institution which uses the software. The increasing presence of ICT for the encoding of metadata has also engendered a shift in the profile of metadata practitioners. In combination with the high volumes of digitized resources to be described, the creation of metadata has become an activity with a more administrative profile than before. Together with the short-term financing of digitization projects leads this to a situation where young, administrative forces work for short
3.6. CONCLUSIONS

periods of time on metadata creation, which has an impact upon the consistency of the metadata quality.

An overview and critical evaluation of the latest emerging approaches of metadata creation and management closes up the chapter. Both ontologies and folksonomies position themselves as alternatives of traditional controlled vocabularies, such as thesauri. The proponents of ontologies claim that, by stating metadata in a highly formalized manner with the help of RDF and OWL, automated reasoning and an ultimate interoperability between resources becomes possible. However, the practical implementation of semantic web technologies is quite problematic, as the workload to prepare the needed metadata is extremely high.

As a result of their uncontrolled character, folksonomies stand completely on the other side of the indexing language spectrum. Loved by policy makers for the apparent democratizing and free character, the actual use of folksonomies is however not so much to be found within the advantage these user-generated classifications offer for effective information retrieval, but more as a communication tool for a cultural institution to facilitate the interaction with their users.
The emergence of the “use-neutral” approach is a promising tendency for certain metadata needs. Confronted with ever evolving user-needs and the pragmatic impetus to re-use existing metadata, mash-ups offer exiting possibilities in certain contexts to build new services upon existing metadata. It should be noted that the quality of the existing metadata plays a crucial role in the success of this approach.
## Chapter 4

### Trends in metadata quality research

<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction .................................................. 102</td>
</tr>
<tr>
<td>4.2 Metadata quality in other application domains ... 102</td>
</tr>
<tr>
<td>4.3 Quality control in library cataloging ................. 106</td>
</tr>
<tr>
<td>4.3.1 Development of copy cataloging .................. 106</td>
</tr>
<tr>
<td>4.3.2 Defining quality in terms of speed and costs of access .......... 107</td>
</tr>
<tr>
<td>4.3.3 Google Books project ................................. 108</td>
</tr>
<tr>
<td>4.4 First steps outside the library community ............. 112</td>
</tr>
<tr>
<td>4.5 Good enough approach: accepting bad quality metadata ......................... 116</td>
</tr>
<tr>
<td>4.6 Conclusions .................................................. 117</td>
</tr>
</tbody>
</table>

**Summary:** This chapter gives an overview of the latest trends in metadata quality research in and outside of the cultural heritage sector. As the biodiversity domain is currently one of the most active fields, some metadata quality tools developed within this discipline are presented. From the library world, the evolving notion of quality cataloging is discussed, by focusing on the notion of the cost/benefits relation. A limited set of automated metadata quality tools designed for cultural heritage metadata is then evaluated and I conclude with the emerging trend of accepting bad quality metadata.
4.1 Introduction

The issue of metadata quality only came to the surface within the cultural heritage domain 10-15 years after the initial research from the business and statistical domain. As long as the metadata of cultural heritage institutions remained within the safe boundaries of the institute the notion of quality was not really an issue. Users passed their question to a member of the staff that could query the database for them. As such, the database (and the metadata records it contained) was more or less considered an internal tool. But then came the web. Initially, most museums limited themselves to an online presentation of their institution. Only a very limited number of museums published their metadata in the same way as libraries, which offered early on their users an Online Public Access (OPAC). The growing tendency to aggregate thematically or geographically related metadata from libraries, archives and museums with the use of OAI-PMH raised the pressure on museums to publish or distribute all of their available metadata. The disappointing level of search results and the minimal descriptions attached to retrieved objects within such projects has finally initiated a discussion on issues such as the consistency, accuracy and completeness of metadata.

This discussion is needed badly, as collection holders increasingly try to re-use legacy metadata and gain more value from them within digitization projects. Metadata practitioners assisting digitization projects that aggregate metadata of different partners must acknowledge that the quality of existing metadata is hardly questioned. After all, which collection holder wants to stand up in the middle of his peers and warn them about the low level of his metadata? However, this misplaced trust causes delays and failures when metadata do not live up to expectations. More importantly, we have to acknowledge that there are no established standards, methodologies or tools for metadata quality evaluation. Or to put it in the often-cited words of Diane Hillmann: “There are no metadata police” [73].

4.2 Metadata quality in other application domains

The area of biodiversity is currently one of the most interesting fields active in the domain of metadata quality. The growing concerns on climate change and global warming of the last decade catalyzed the international research on biodiversity. This research community is particularly interesting for the cultural heritage sector, as both domains share a mutual concern of interoperability and the exchange of metadata. Large international research projects such as the European Network for Biodiversity
4.2. METADATA QUALITY IN OTHER APPLICATION DOMAINS

Information (ENBI) aggregate huge sets of scientific data concerning the climate, flora and fauna that were originally created and maintained by local agencies.\(^1\)

The success of the aggregation of metadata from diverse sources depends on the quality of the individual repositories. The Reference Center for Environmental Information of Brazil therefore developed a data cleaning tool, which helps curators to identify possible errors.\(^2\) Figure 4.1 presents a screenshot of how the database presents suspect records.\(^3\) The system passes recommendations to the initial author or curator of the metadata.

![Data cleaning tool from the biodiversity domain](image)

Figure 4.1: Data cleaning tool from the biodiversity domain

---

1. See http://www.enbi.info
2. See http://www.cria.org.br
3. The tool is available on http://splink.cria.org.br/dc
CHAPTER 4. TRENDS IN METADATA QUALITY RESEARCH

geo-referenced, how many duplicated records have been detected, when the last update of the collection took place, etc. Each time suspect records are mentioned, a direct link is provided to verify manually in detail the record and its metadata.

Amongst the options offered on this page, I especially would like to point out the possibility to visualize the data cleaning statistics, which offer graphs representing the evolution through time of the number of suspect authors, duplicated records and catalog numbers. This tool, presented in Figure 4.2 offers the opportunity for a potential user of the collection to grasp within 10-15 minutes the quality of the data he or she is interested in.

Figure 4.2: Graphs representing the evolution of data quality within the biodiversity domain

Another interesting example from the biodiversity domain comes from the Integrated Taxonomic Information System (ITIS) database, developed and maintained by the National Museum of Natural History in Washington. This database has been build up toward the end of the 1990's with legacy metadata, which has varying levels of quality ranging from one taxonomic
4.2. METADATA QUALITY IN OTHER APPLICATION DOMAINS

Errors range from typographical errors to species names that have no author or date, or that have been assigned to the wrong group. From the initial legacy dataset consisting of 210,000 records, ITIS has grown to nearly 478,000 scientific names, from which 77% have been verified in the meanwhile.

The verification process includes three data quality indicators:

- **Record credibility rating**: states whether the records has undergone internal review or not. When the legacy metadata were imported, they received a record credibility rating of unverified. As ITIS is reviewing the records, the rating changes to either the highest value, verified standards met, when the information within a record is backed up with peer-reviewed or academic references, or the middle value of verified minimum standards met when the record has been reviewed but considered incomplete and/or containing some errors, or draws upon information from a non-peer reviewed source.

- **Latest record review**: this represents the year that the record was last reviewed by ITIS. As taxonomic metadata evolve quite often, this is important information for researchers.

- **Global species completeness**: indicates whether all known valid or accepted species for that rank are incorporated into ITIS at the time of review.

The Canadian partner-developed website that contains a periodically mirrored ITIS data set offers an extra feature for reviewing the status and level of verification of taxonomic groups in ITIS by using of color-coding to represent the Record Credibility Rating for given name records. When a name is mentioned in a report, a link is provided to a table (see Figure 4.3 for an example) showing the number of valid/accepted subordinate taxa, by rank and level of review.4

The most important feature of these tools is their ability to monitor the evolution of metadata quality through time. As the evaluation is completely automated, it can be processed periodically without any supplementary costs, which allows to generate reports on the evolution throughout time of metadata quality, as illustrated in Figure 4.2. The iterative character of metadata quality evaluation is an aspect that will be incorporated within the third strategy presented in Chapter 7.

---

4The tool is available on http://www.itis.gov/itis_primer.html.
CHAPTER 4. TRENDS IN METADATA QUALITY RESEARCH

The discussion regarding metadata quality only appeared quite recently within the cultural heritage sector. However, as in many other contexts, the library sector has been a forerunner. The work on bibliographic control within large national libraries obliged catalogers to reflect on the notion of quality. In her article on the history of quality within bibliographic catalogs, Sarah Thomas describes how Andrew Osborn of the Library of Congress reports in 1941 on the so-called “crisis in cataloging” that arose from the cumbersome way of producing high quality bibliographic records. This already led some 70 years ago to a tension between the ambition to produce accurate and complete records and the reality of having only limited resources at hand:

The word "quality" does not enter into Osborn's discussion of the so-called crisis in cataloging. Rather, Osborn employs the concept of perfectionism to describe a hyper-emphasis on exactitude and precision - in this case, quality gone awry by being taken to an extreme. That today's problems so closely
mirror those of a half-century ago, despite dramatic technical advances and a growing body of literature about library patrons and catalog users, gives one pause [141, p. 495].

Thomas continues by demonstrating that the interpretation of quality within the bibliographic control domain evolved throughout time:

At the Library of Congress, an institution which is surely seen by most, if not all, practitioners of cataloging as the sine qua non for bibliographic control, every record is deemed a sterling example of “quality cataloging” - i.e., every full original cataloging record. For the past decade, the standard for quality cataloging has been altered, some would say eroded, by compromises that include tolerance for, and even the embracing of, such cataloging practices as minimal-level cataloging, collection-level cataloging, copy cataloging and, most recently, the core bibliographic record. To review the changing values in this area of bibliographic control, it is useful to examine these cataloging variations and their relationship to a corresponding shift in the definition of quality associated with them. Quality is not immutable but is rather a standard of excellence that reflects the values of the individuals proclaiming it [141, p. 496].

The shift in the definition of quality is reflected within the Library of Congresses’ tradition of creating full and original cataloging records, which were from 1901 on distributed to other libraries in the US and worldwide. Huge investments were made to assure the highest possible quality, which was ensured by strict rules and several layers of control before a record was submitted to the library community.

It is also crucial to acknowledge the evolving character of how metadata quality itself is perceived or defined. Before the practice of copy cataloging existed, catalogers concentrated on the internal quality of their catalog, as no outside sources were being used to enrich their catalog. With the growing practice of copy cataloging, more and more importance was given to how the quality of in-house records related to the quality of records externally brought in. More importantly, within the context of a card catalog which could only be accessed locally by one user at a time, the idea of striving toward a perfect, flawless catalog was feasible. The rather limited access to the catalog considerably lowered the number of errors discovered within the catalog [75].

4.3.2 Defining quality in terms of speed and costs of access

Gradually, libraries were confronted with backlogs as a result of the important delays between the time of publication of the books and the avail-
ability of the catalog record from the Library of Congress. Confronted with this problem, libraries increasingly started copy cataloging from other sources, which meant in some cases lowering the high standards set by the Library of Congress. This evolution points out an important shift in the notion of quality, in the sense that the understanding of metadata quality no longer was entirely based on the notions of accuracy and completeness, but started to include the more service-oriented element of timeliness. Users need to access a new publication as quickly as possible, even if it means compromising the quality of the cataloging. The first purpose of a library is to satisfy the needs of its readers by offering a catalog which is “comprehensive, current and usable but not necessarily detailed or even perfectly accurate” [102].

In parallel with this more user-centered approach, the call for a more cost-effective and productive cataloging was increasingly heard in libraries throughout the 1980s and the 1990s. Cost savers proclaimed that too much money was spent to live up to the high norms for the attribution of subject headings, all the more since users have difficulties using them. Within this general atmosphere of cost-savings, the Library of Congress launched the practice of the “core” record, which took 20% less time to create by focusing only on the essential elements of the record.

The quest for more productivity also introduced the outsourcing of cataloging and other library services. The subcontracting of activities gradually became a common practice in every possible sector at the beginning of the 1990s, but mostly only non-core competencies of organizations, such as catering and janitor services, are subcontracted to a private company. Outsourcing in the library sector is in that regard somewhat peculiar, in the sense that cataloging can hardly be considered not to be a core activity or mission of a library. Roger Martin wrote a report on the impact of outsourcing and privatization on library services and management, which was commissioned by the American Library Association, in which he concludes that “in the main outsourcing has been an effective managerial tool, and when used carefully and judiciously it has resulted in enhanced library services and improved library management” [99]. But there is not really a consensus on this matter within the sector. Basically, adherents point out lowered costs and a drastic increase in the speed with which users can have access to new publications. Critics of outsourcing say the level of the subject indexing is lower and that it threatens the inhouse competencies [138].

### 4.3.3 Google Books project

The discussion on what quality cataloging exactly is, and the ways cataloging standards and rules should correspond with the limited resources of libraries and evolving user needs, now enters a new stage with the im-
4.3. QUALITY CONTROL IN LIBRARY CATALOGING

Implementation of the Google Books project. The importance of the impact of the Google Books project for the future perception of quality cataloging within libraries should not be underestimated. The University Library of Ghent is currently collaborating with Google to digitize the books from the library that no longer fall under copyright restrictions.

It is interesting to study the different transitions that the library catalog has been through. By taking a concrete example of the catalog record of a specific book, the evolutions in the metadata can be analyzed.

The library started creating paper-based catalog cards in 1912, which resulted in a systematic catalog in 1921. Figure 4.4 represents a catalog card as it was created during that period.

In 1984-85 the first catalog database, under the name of Aleph, was developed. The metadata from the paper catalog cards were gradually converted by hand to Aleph as library users consulted them. In 2004-2005 a large scale scanning operation resulted in a digitization of all the catalog cards, which were then made available in the CatFich database.

With the help of OCR techniques, the content of the scans of the paper catalog cards was converted to structured textual documents, but errors were inevitable. In the context of the Google Books project, the scans of the paper cards are currently used to check whether the OCR version of the metadata is correct. Figure 4.5 illustrates the back-end of the Aleph database. One can notice that a great number of extra metadata appeared

---

5For more background on the Google Books project, see http://books.google.com/
6The background information on the library catalog was provided by members of the cataloging department of the library.
7This database is available on http://search.ugent.be/catfich-browse/
in comparison with the paper catalog file, such as for example the Dewey categories. The original catalog file has also been enriched with existing metadata from the LIBIS network, which is indicated by the Belvls code included within the 040 field.

![Screenshot of the Aleph database back-end](image)

Figure 4.5: Screenshot of the Aleph database back-end, which also incorporates a scan of the card catalog represented in Figure 4.4.

It is interesting to confront the gradual enrichment of the catalog, which was very resource intensive, with the way Google is currently disclosing the books through their Google Books interface. The library has provided an export of the existing metadata records from the Aleph database, but it is unclear whether Google actually uses these metadata, such as the Dewey categories. Google unfortunately does not want to communicate on whether, and if so, how they re-use the existing metadata delivered by participating libraries.⁸

Figure 4.6 represents the interface in which the metadata of the book, whose metadata record was also represented in Figures 4.4 and 4.5, are

---

⁸I have emailed with the associate product manager of the Google Books project, but I did not receive any answers regarding their approach toward the re-use of the existing metadata.
4.3. QUALITY CONTROL IN LIBRARY CATALOGING

displayed. In this interface, the traditional metadata such as the title, author, place and date of publication only take up a limited part of the search and browse possibilities. Even taking into account that the fact that the back-end of the Aleph database and the front-end application of Google Books are two different types of interfaces, it is fascinating to see that automatically generated metadata, such as extracted key words, phrases and paragraphs which are quoted in other books, offer the most possibilities of browsing and retrieving information within the Google Books interface.

Figure 4.6: Screenshot of the Google Books interface presenting metadata of the same book as mentioned in Figures 4.4 and 4.5

The arrival of Google Books is an important event within the history of library cataloging and the evolution of the concept of quality cataloging. The project is still at a beta-stage, but the results and possibilities offered by the Google Books interface seem to be promising. If the company can deliver in the long run high end-user satisfaction by relying on brute computer power and natural language processing, then this would definitely give another blow to the traditional community of catalogers who criticize the down-grading of professional cataloging.
4.4 First steps outside the library community

Bruce and Hillmann provided the first major conceptual framework of metadata quality for the cultural heritage sector [28]. Defining quality measurements and frameworks is important, but they also have to be put into practice.

The manual analysis of samples of metadata records has been the most applied method to gather metadata quality indications [131]. However, this manual approach has two obvious disadvantages: 1) it is time consuming (and thus expensive) and 2) it only offers a “snapshot” of a sample of the metadata records at one specific moment in time. Therefore, I will focus only on semi-automated approaches which analyze the totality of a given metadata set.

Tennant proposes a minimal, but pragmatic set of analysis functions to be applied on metadata and specifies queries to be computed such as the total number of occurrences of a certain value or patterns across records (e.g. all records with x in the y field do not have a z field) [146]. The application of such scripts or queries on large numbers of metadata records evidently produces results which are difficult to grasp without the aid of visualization software. Dushay and Hillmann present a tool that can translate the results of queries upon a large collection of records into a human-readable format that allows the detecting of patterns and the extent of the problems [46].

Several researchers have also worked on metadata transformation and enrichment, especially in the context of aggregated content projects. Foulonneau and Cole report, for example, on how harvested records can be transformed to be of higher use in the context of an OAI service provider [55]. A research team from the University of Waikato in New Zealand launched a tool that generates statistics and visualizations of OAI repositories. Figures 4.7 and 4.8 illustrate the type of visualizations and the statistics the tool delivers [106].

The automated metadata quality assessment methods we have seen so far focus on the formal characteristics of metadata, and not on its actual content. The evaluation of the accuracy and conformance to expectations of metadata is left to human evaluation. Ochoa and Duval however propose to translate these and the other criteria from the Bruce and Hillmann framework into equations that can be automatically applied [109]. This approach has already been criticized in 1.3.1.1. Their method is based upon a comparison between the actual content of the metadata record and the resource itself. The approach of detecting how many words match between the metadata record and the resource it describes can first of all only function for machine-readable textual documents. Moreover, the idea that the content of a metadata record literally reproduces the content of a resource is a very reductionist vision on the methods and goals of indexing.
4.4. FIRST STEPS OUTSIDE THE LIBRARY COMMUNITY

A big step toward the assessment of metadata quality was taken with the development of metadata application profiles, as their goal is to explicitly communicate to a user community what to expect from the metadata they are consulting. Heery and Patel introduced the concept of application pro-
files in 2000 [73]. The large gap that exists between the standard makers and the implementors acted as a starting point for their reflexion. The developers of schema are concerned by the coherence of an element set, that should be in theory be able to be used by all of the users of the community the standard addresses. The main interest of the implementors is to effectively describe their own, specific resources, which means that standards are in most cases adapted to local practices and needs. Application profiles recognize the inevitable need to localize or tailor standard metadata schemes, by incorporating “data elements drawn from one or more namespace schemas combined together by implementers and optimized for a particular local application” [73].

The idea fostered in the next few years. The European Committee on Standardization (CEN) financed the development of technical specifications, but the effective use of application profiles has been mainly promoted by Dublin Core communities, such as the DCMI Libraries Community and the DCMI Collection Description Community, which developed specific application profiles. Figure 4.9 represent a fragment of a Dublin Core Collection Application Profile [76].

So what is the link with metadata quality? Application profiles offer a way to inform users of metadata on what they can specifically expect from those metadata. Hillmann and Phipps adequately call them “templates for expectations” [76]. The idea of managing the expectations of users from metadata or information retrieval systems has also been applied within other domains. The INSPEC bibliographic database of engineering literature has attached to each record a specific indication of the “expertise level”, in order to let the different user communities identify which information suits their needs [2].

As long as this information is only accessible in a human readable form as html-documents, the use of application profiles is quite limited. However, unleashing the real power of application profiles by automating their validation is not easy. As Hillmann and Phipps rightly point out, the validation of RDF, that underlies the “Open World” assumption, is not possible. The problems to implement automated or machine-readable application profiles point out in a very clear way to the recurring problem between the enormous advantages the semantic web could offer, but on the other hand it’s inability to be applied in day-to-day practice:

Further complicating the issue for data consumers seeking RDF data validation is RDFs reliance on the assumption of an “Open World” of data in which data that is not present in a description set cannot actually be considered to be missing, but simply is not able to be dereferenced or retrieved at the present time. “Missing” data may also not be considered missing if its existence can be inferred based on conditions provided by an OWL or RDFS definition. These assumptions are highly useful when
4.4. FIRST STEPS OUTSIDE THE LIBRARY COMMUNITY

This brings us to the way the validation of the syntax of a document corresponds to notions of metadata quality. The header within a HTML document states which version of HTML is respected. This can then be checked by using the W3C validator to see if no rules were broken with the HTML tags. The same applies with XML documents that specify a DTD, so that XML files can be checked. The case of application profiles implemented in RDF is more complicated, as Hillmann explains by referring to the problems of the “Open World” character of RDF.

Figure 4.9: Illustration of a human-readable fragment from the Dublin Core Collection Application Profile
4.5 Good enough approach: accepting bad quality metadata

The previous sections illustrate the huge efforts that have been made in the last decades to create metadata in a standardized and shareable manner. The investments needed before some tangible results can be presented are so high, that more and more metadata practitioners are adopting a very pragmatic approach. An increasing number of collection holders publish their metadata, knowing that they might be of bad quality.

Traditionally, cultural heritage workers have a long standing custom of publishing highly authored and verified content. This mentality has been largely transposed to the sphere of the Internet and online publication. It is only quite recently that the younger generation of metadata practitioners have been advocating a much more pragmatic way of approaching the publication of metadata. Under the motto “Better a bad access than no access at all”, institutions are more and more considering their collection description as an ongoing work in progress.

Some heritage institutions are effectively stating that their metadata quality is bad and call in the help of the users to augment the metadata. The following quote was extracted from the website of the image database of the National Archives of the Netherlands:

> The quality of the descriptions is not optimal. Therefore, it can happen that the results of a query return less images than there are actually are within the database. The National Archives are working very hard to augment the quality of the descriptions, so that you will obtain even better results in the future. If you find an error in the description of a photograph or if you would like to add some information to it, you have the possibility to send us your remarks by using the “comment” link.9

The Powerhouse Museum displays a somehow similar statement on their website with relation to the quality and the completeness of their collection registration database:

> Most of our collection has been acquired before the advent of computers and as such a large portion of Museum knowledge and documentation is held in paper and non-digital formats. The level of documentation varies depending on when the objects were collected, what was known about them at the time, the research that has been undertaken since, and whether this research has been digitised. You will notice that more recent acquisitions are more likely to have extensive documentation available online, and often, better quality, colour, zoomable images.

---

9See http://beeldbank.nationaalarchief.nl/
4.6. CONCLUSIONS

Research and documentation, both textual and photographic, of the Museum’s collection is a continuous process.¹⁰

These statements appear quite logical and obvious, but they do reflect a crucial change in the mentality of heritage institutions. Until recently, the high editorial standards that accompanied the different types of publications, such as exhibition catalogs, were transposed to the world of online publication. Only gradually rose the idea that online publication is an entirely new way of giving access to information. The preservation of the high standards blocks out the new possibilities offered by the Internet, such as the speed of publication and the interactivity with users.

In analogy with the “Open World” versus the “Closed World” assumptions from the database world, one could state that we move from an “Ideal World” to “Messy World”. We have evolved from a total quality perception to a situation where the costs and timeliness of metadata are also taken into account.

4.6 Conclusions

The review of some metadata quality tools from the biodiversity domain in Section 4.2 has demonstrated the importance of not only producing a “snapshot” of the quality of metadata records, but also the importance of performing analyses on a regular basis, which allow to study the evolution of metadata quality throughout time.

Not only the quality of metadata evolves, the definition and interpretation of metadata quality itself is not a fixed concept. This is the major conclusion that can be drawn from the evolution the concept of “quality cataloging” has known throughout the history of library cataloging.

Starting from the idea of the absolute perfect catalog presented in Section 4.3.1, a more pragmatic approach was adopted throughout the years as catalogs were increasingly opened up and shared. Economic imperatives also obliged libraries to cut down on the high standards and even to outsource cataloging to private companies. Section 4.3.2 presented how the timely delivery of new catalog records toward users is also increasingly considered a crucial element for user satisfaction.

The tension between the manual, labour-intensive cataloging and the more pragmatic approach that recently has been taking the lead, was demonstrated in Section 4.3.3 with the help of a catalog record from the University Library of Ghent. Three stages (paper catalog card, record within the back-end of the Aleph catalog and the Google books interface) allowed to follow the enrichment of the metadata contained on the paper catalog card within the Aleph database. The inclusion of automatically

¹⁰See http://www.powerhousemuseum.com/collection/database/
created metadata within the Google Books interface questions the validity of the manually created metadata.

The last five years, a limited number of metadata quality tools specifically aimed toward the cultural heritage community have been developed. Most of the tools presented in Section 4.4 concentrate on the “completeness” aspect of metadata quality, by indicating which metadata fields actually have content. The work on metadata application profiles offers interesting perspectives, by the way they offer the opportunity to communicate on the level of expectations that users can have when they consult metadata from a specific collection. The automated validation of application profiles is currently however not operational.

Section 4.5 points to what is identified as the most important shift in the thinking on metadata quality within my application domain. After several years of projects that did not manage to deliver metadata that met pre-Internet quality standards within reasonable amounts of time, a more radically pragmatic approach is emerging. As quotes from the websites of two representative heritage institutions demonstrate, cultural heritage institutions increasingly publish metadata that do not meet quality standards under the motto “Better a bad access than no access at all”. So finally, here again, we are confronted with the increasing subordination of metadata quality criteria in face of user needs.
Part II

Dealing with metadata quality in practice: three original strategies
Outline of the second part

Having sketched the necessary historical and technical background of metadata quality in the first part of the dissertation, this second part will develop strategies on how cultural heritage institutions can act upon their metadata quality in a pragmatic and a low-cost manner. The different strategies will also demonstrate in a direct way how the daily interaction with metadata themselves, analyzed within the “temps court”, can influence evolutions from the “temps intermédiaire” and the “temps long”. The second part of the dissertation therefore acknowledges the two-way interaction between the different time levels from the stratified time concept which was theorized in Section 1.2.3.

The first step of the threefold approach consists of analyzing the existing metadata quality, as no other action can be undertaken unless there is a report on the status of the metadata. The methodology of data-profiling allows us to obtain statistical information on the formal quality of metadata. The application of a set of five simple analyses upon metadata from the Royal Museum of Central Africa will demonstrate how this kind of strategic information can be obtained in a quick and straightforward manner.

After the analysis of the formal characteristics of metadata, some considerations are needed as to how an institution can act upon bad quality metadata. As the previous step can only take into consideration the formal quality of metadata, and the manual correction of metadata by professionals is too expensive, I propose to investigate the possibilities offered by user comments to augment the quality of metadata within cultural heritage databases. Two maximum variation case studies deliver empirical data on the relevancy of user comments. The first case study consists of the image database of the National Archives of the Netherlands, which offers a huge amount of historical photographs that appeal to the general public. The Ross Archive of African Images offers the second case study, and differs from the first case study by its highly specific content and closed access to a limited number of domain experts.

The third and last approach introduces the idea of pro-actively monitoring the concrete metadata needs of users. This can be done by offering a dynamic search interface in the database, where users can configure metadata fields in a very intuitive manner to customize the search interface to their personal needs. The interaction between the users and the search interface delivers statistical information on how and which metadata fields are being used by users and which are not. This information, obtained from the daily interaction between users and metadata from the short time, can then be used to guide future actions upon the metadata scheme, studied in the intermediate time, and influence the development of a long-term metadata management strategy.
Chapter 5

Applying data-profiling upon metadata records

Contents

5.1 Introduction ............................................ 124
5.2 Background of data-profiling in other domains .... 125
5.3 Data-profiling within the cultural heritage sector .... 126
5.4 Case study: the Royal Museum of Central Africa .. 127
  5.4.1 Presentation of the analyzed metadata set and its context .............................. 128
  5.4.2 Presentation of the data-profiler tool ........................................... 131
  5.4.3 Results of the analyses .................................................. 131
5.5 Conclusions ............................................... 138

Summary: This chapter presents data-profiling as the first step toward metadata quality, by offering a methodology to analyze the formal quality of large metadata sets. The initial context of use within the business and administrative domain, and the originality of applying data-profiling within the cultural heritage domain is outlined in the first sections. Central within this chapter stands the application of a general purpose data-profiling tool upon a large metadata set from the Royal Museum of Central Africa. The conclusions link the results of the analyses back to the context of the metadata creation.
5.1 Introduction

Metadata creation and management is a continuous and iterative process. Cultural heritage institutions not only need to manage the incessant arrival of new objects, their catalogs and databases also reflect the history of the different methods and tools used for the description of their holdings. As Chapter 3.2 demonstrated, the succession of technologies is marching at a considerably faster pace the last decade then ever before. Collection description software currently has a general lifespan of three to six years, but this lifespan is becoming increasingly shorter.

Not only the tools succeed one another, the practitioners who use the tools for metadata creation and management also change at a quicker pace then in the past. With the prevailing project-based way of managing digitization projects, indexers and catalogers no longer work for several years on the same collection, but switch very often from one project to another. Section 3.4.2 has already pointed out to this issue. Research on inter-indexing consistency has showed that significant differences occur amongst professional catalogers and indexers occur, but the differences can be assumed to be a lot higher when young people with no formal training in metadata creation take up this activity for a short period of time [81].

This permanent succession of both the technologies and the professionals concerned with metadata creation and management has led to a situation where more and more incoherences slip into collection registration databases. When confronted with a database that has ten- or hundreds of thousands of records, it becomes impossible to obtain an impression of the overall quality of the metadata records with manual methods. Statistical significant samples can be checked manually, but this method is very resource intensive.

It is exactly in this context of uncertainty and the need for automated solutions that I wish to draw upon the methodology and tools known under the term of data-profiling. Jack Olson defines data-profiling as the use of analytical techniques to discover the true structure, content, and quality of a collection of data [110]. The following chapter will introduce its functionalities and elaborate on why this methodology from the business and administrative domain is of interest to my specific application domain. A case study from the ethnographic department of the Royal Museum of Central-Africa (RMCA) will illustrate which type of errors can be found when applying data-profiling upon a large set of museum metadata.

---

1 I published an article which resumes the findings of this chapter [151].
5.2 Background of data-profiling in other domains

As I already mentioned in the first chapter when defining metadata quality, the business domain gradually acknowledged in the beginning of the 1990s the effect of bad quality data upon business profits. In parallel, academic research and consultancy activities developed around the issue of data quality. The book “Data quality: the accuracy dimension” of Jack Olson is one of the major works which present a structured overview of the quality assessment of databases and information systems [110].

Olson identifies data-profiling as a core activity within data quality research, as it necessarily is the starting point for any data quality research. In contradiction to data-mining, which is another technique to analyze large sets of data, data-profiling concentrates on the instance analysis of individual attributes, from which it extracts information such as data type, length, value range, discrete values and their frequency, variance, uniqueness, occurrences of null values and typical string patterns [121].

The goal of data-profiling is to detect data quality issues. But what can the different problems be? Rahm and Hai Do provide a typology of data quality problems. On a first level, they distinguish single source from multiple source problems. As I will mention in the next section, I am not interested in the problems that appear in aggregated data, as some research has already been done on the quality of aggregated metadata in cultural heritage portals [55].

Therefore I only look at the problems that occur with single source data. Within this subgroup, a crucial difference needs to be made between problems at instance and schema level. Problems that occur on the schema level are of course reflected on the instance level, but instance-specific problems “relate to errors and inconsistencies that cannot be prevented at the schema-level” [121]. Figures 5.1 and 5.2 from the article of Rahm and Hai Do resume examples of the difference between the two types of problems [121]. Data-profiling intervenes in these problems by creating in an automated way metadata that can help us detect the above mentioned problems. Figure 5.3 gives an overview of the type of metadata that can be created with data-profiling to identify data quality issues.

<table>
<thead>
<tr>
<th>Scope/Problem</th>
<th>Dirty Data</th>
<th>Reasons/Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>val=001.55.70</td>
<td>values outside of domain range</td>
</tr>
<tr>
<td>Record</td>
<td>age=22, bdate=12.02.70</td>
<td>age = (current date - birth date) should hold</td>
</tr>
<tr>
<td>Record type</td>
<td>emp.c=(name=&quot;John Smith&quot;, SSN=1234567)</td>
<td>uniqueness for SSN (social security number) violated</td>
</tr>
<tr>
<td>Source</td>
<td>emp.c=(name=&quot;Peter Miller&quot;, SSN=1234567)</td>
<td>uniqueness for SSN (social security number) violated</td>
</tr>
<tr>
<td>Source</td>
<td>emp.c=(name=&quot;John Smith&quot;, deptno=127)</td>
<td>referenced department (127) not defined</td>
</tr>
</tbody>
</table>

Figure 5.1: Examples of problematic data at schema level
CHAPTER 5. APPLYING DATA-PROFILING UPON METADATA RECORDS

Figure 5.2: Examples of problematic data at instance level

Figure 5.3: Examples of metadata created by data-profiling to detect data errors

5.3 Data-profiling within the cultural heritage sector

The cultural heritage sector is slowly starting to realize that it has to act upon the quality of its metadata. Section 4.4 already gave an overview of the research that has been conducted on metadata quality in the sector, which consists mainly of theoretical work on metadata quality frameworks, such as that of Bruce and Hillmann [28] and the assessment of library metadata in an aggregated context, such as the work of Foulonneau [55, 54]. The automated tools that were presented (see Figures 4.7 and 4.8) focused largely on the “completeness” aspect of metadata, but again in the context of aggregated metadata.

Little or no research has been done to develop methodologies and tools for checking metadata quality within single source cultural heritage databases.
outside the library community. To my knowledge, I am the first to apply data-profiling to cultural heritage collections. Closest to my research is the metadata analysis tool that Mark Philips from the University of Texas library has created in 2007. The tool is not documented in an article or in some other publication, but is demonstrated in an online presentation. The tool checks which Dublin Core metadata fields are used by different partners within a digital heritage repository and visualizes when metadata records are created. But again, the emphasis is on the multi-source aggregation of metadata.

During an interview with Peter Schouten, who is a consultant for the Dutch private enterprise Ingressus, he demonstrated me how his enterprise applies in a limited way data-profiling techniques when assessing metadata for customers. The company has developed a custom tool, unavailable for third parties, which can analyze metadata in a similar way as the data-profiler presented in the case study of this chapter. The Ingressus tool has a more limited set of scripts. Figures 5.4 and 5.5 demonstrate a bibliographic record which is analyzed by their tool and the resulting report. The consultants of Ingressus use the report as a starting point for metadata improvement procedures and consultancy reports on the quality of the customers' metadata.

The scripts are based on the presence of null-values, the length of the values of fields and the number of times a specific field appears within a record. No pattern analysis is available. This tool is only available to customers of Ingressus, who are solely confronted with the end report on the status of their metadata. I claim that the technology behind data-profiling is straightforward enough to be deployed on a much larger scale than it is currently done. Cultural heritage institutions should be able to perform the analyses on their own. They badly need a methodology and a seamless integration of data-profiling tools within their collection registration software.

5.4 Case study: the Royal Museum of Central Africa

This section will present the test collection of metadata and their context, and the general purpose data-profiling tool that will be applied on these metadata records.

---

3 Ingressus provides metadata creation and clean-up services toward libraries and other heritage institutions. See the company's website on http://www.ingressus.nl.
CHAPTER 5. APPLYING DATA-PROFILING UPON METADATA RECORDS

Figure 5.4: Illustration of the custom built data-profiler of Ingressus on a collection of bibliographic records

5.4.1 Presentation of the analyzed metadata set and its context

The utility of the data-profiling is demonstrated by using an export of the collection registration database of the ethnographic department of the Royal Museum for Central Africa (RMCA). The museum is one of the ten federal scientific establishments in Belgium and functions at the same time as a scientific research institute dedicated to Africa and a cultural center.4

4See the website of the museum on http://www.africamuseum.be.
5.4. CASE STUDY: THE ROYAL MUSEUM OF CENTRAL AFRICA

The cataloging and indexing of the collection of ethnographic already started before the actual founding of the museum in 1910, with the use of paper catalog sheets which contained the pre-printed fields “region”, “number”, “group”, “description”, “origin”, “use” and “remarks”.5

A second version of the catalog was initiated in 1910, and contained extra fields as “classification”, “tribe”, “district”, “area”, “location”, “donator”, “administration” and “ethnical notes”. The field “classification” and the more detailed geographical information reflects the development of the colonial bureaucracy and ethnography after 1910.

A third and last version of the paper catalog was started shortly after

---

5The history and context of the development of the database from the ethnographic department is based upon the article of van Hoolland and Vanhee [132].
the Second World war. New elements were the differentiation between “donation”, “loan” and “purchase”, the inclusion of bibliographic fields related to publications and fields with more information to the treatment, conservation and location of the object within the collection. These new fields illustrate the expanding character of the collection and the growing need to professionalize the conservation of the objects. This catalog was expanded and consulted until the early years of the 21st century.

In parallel with the paper catalog, a first electronic version was created in the 1990s to prepare the large-scale moving of the ethnographic collection to another location. By using the state of the art technology of that period, dBase III, a database was designed as an internal working instrument to master the practical management of the collection. For pragmatic reasons, only seven fields were included from the paper based catalog in the database: “inventory number”, “description”, “ethnical background”, “region”, “country”, “location” and “location number”. The database was thus a semantically poorer version of the paper-based catalog, as the database was merely considered a practical instrument to follow the objects before and after the moving of the collection. Most scientists continued to use the paper-based catalog and no public access of the database was envisioned.

The relocation of the collection obliged the collection registration team to describe in a quick and pragmatical way a large number of objects, which led to descriptions that were simply based upon formal analogies, resulting in descriptions such as “fly” or “sausage”. The encoding of the ethnical and geographical metadata also took place in the urgency of the moment, which led to a very unsystematic metadata encoding. Figure 5.6 illustrates the extent of the problem with an overview of the 47 different ways of encoding the name of the ethnographic group “Chokwe”.

From 2000 onwards, the department realized that a collection registration software was needed to replace the rigid and user-unfriendly dBase III application. After a long period of comparing different applications, The Museum System (TMS) was acquired. The purchase of this software was accompanied by growing expectations regarding the online publication of the metadata but also of digitized images from the ethnographic collection. However, after a short period of initial enthusiasm, it became clear that the migration of the legacy metadata from dBase to TMS was a tedious task and, more importantly, that the existing metadata needed to be upgraded.

The database manager of the ethnographic department created in September 2007 a .csv export file from the SQL Server database back-end on which the collection registration software runs, consisting of 69,719 records. Each record has 13 fields (object id, object number, object count, date of collection, date of entry, date of production, title, medium, dimensions, thesaurus terms, description, old region, actual region). The metadata are written in French, with some minor exceptions of Dutch.
5.4. CASE STUDY: THE ROYAL MUSEUM OF CENTRAL AFRICA

5.4.2 Presentation of the data-profiler tool

The analyses upon the .csv files of the RMCA are performed by an open source general purpose data-profiling tool, developed by Yves Bontemps and freely available as open source software.\(^6\)

The data-profiler works in three steps. First, the analysis to perform on the dataset has to be set up by creating an XML profile specification file in which is specified which analysis runs on which column of the dataset. Figure 5.7 presents a screenshot from this specification file. Five analyses are available, which will be presented with the help of examples from the test collection. In a second step, the profiler itself is launched, which will read the .xml file and store the result of the profiling into a local repository and the information about the profiling execution into a catalog file. The catalog file is used to record what profile specification (.xml file) was used as a basis for profiling and to retrieve the results from the local repository. Finally, the visualizer has to be executed to view the profile execution results. Figure 5.8 demonstrates the graphical interface representing the results of the analyses. These can be exported for further analysis in another tool.

5.4.3 Results of the analyses

The different analyses will now be illustrated with examples from the test collection.

\(^6\)The software can be downloaded from the open source software repository Sourceforge on http://sourceforge.net/projects/dataprofiler/.
CHAPTER 5. APPLYING DATA-PROFILING UPON METADATA RECORDS

The nullcount analysis calculates the number of records where the specified column holds no value. Table 5.1 represents the results of the analysis on each available field and illustrates the high number of fields which remain without values. Several fields, such as description, dimensions, date of production, date of collecting and creditline, have no value in approximately 90% of the metadata records, which is quite disturbing.

The calculation of the completeness of a metadata field is very straightforward from a technical point of view, but as Ochoa and Duval note, the outcome does not necessarily correspond with how humans evaluate the completeness [109]. Certain metadata fields are not always as relevant for all types of documents or objects that are being described with the same metadata scheme. Especially fields like "description", which give the broader context of the object, are normally frequently consulted by users.

5.4.3.2 Pattern analysis

The pattern analysis calculates the different formats used to represent values. The values can be alphabetical characters (represented by the profiler with A), numerical characters (represented by the profiler with 9) or other special signs such as a punctuation sign or a slash. This analysis is particularly useful to examine the values which correspond (or should correspond) to a certain fixed syntax, such as accession numbers and dates. The accession number in the case of my data set has to correspond to...
the following fixed syntax: [collection code].[inscription year].[lot number].[number of the item within a lot].[number that indicates that the item is a part of serie]. When running the pattern analyzer, we can see that 92% of the values match the required syntax.

The different date fields also offer an excellent opportunity to apply the pattern analysis. In total, 52 different ways of encoding the date of collecting exist within the database. This is due to the fact that sometimes other information is also saved within the field. Obviously, this should
CHAPTER 5. APPLYING DATA-PROFILING UPON METADATA RECORDS

<table>
<thead>
<tr>
<th>Fieldname</th>
<th>Percentage of empty fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>objectid</td>
<td>0%</td>
</tr>
<tr>
<td>objectnumber</td>
<td>0%</td>
</tr>
<tr>
<td>objectcount</td>
<td>0%</td>
</tr>
<tr>
<td>date of collecting</td>
<td>87.5%</td>
</tr>
<tr>
<td>date of entry</td>
<td>55.6%</td>
</tr>
<tr>
<td>date of production</td>
<td>92%</td>
</tr>
<tr>
<td>title</td>
<td>8%</td>
</tr>
<tr>
<td>medium</td>
<td>66.3%</td>
</tr>
<tr>
<td>dimensions</td>
<td>90.7%</td>
</tr>
<tr>
<td>creditline</td>
<td>89.5%</td>
</tr>
<tr>
<td>description</td>
<td>92.7%</td>
</tr>
<tr>
<td>region old</td>
<td>44%</td>
</tr>
<tr>
<td>region new</td>
<td>44%</td>
</tr>
</tbody>
</table>

Table 5.1: Percentage of empty fields

be avoided. Table 5.2 represents the 10 most frequent patterns used to represent the date when an item was acquired. This clearly demonstrates the need to standardize the input of dates.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Number of occurrences</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(empty)</td>
<td>65,011</td>
<td>1891-1914</td>
</tr>
<tr>
<td>9999</td>
<td>1,564</td>
<td>1909</td>
</tr>
<tr>
<td>99-99/9999</td>
<td>1,105</td>
<td>09-10/1992</td>
</tr>
<tr>
<td>99/9999</td>
<td>574</td>
<td>01/1994</td>
</tr>
<tr>
<td>99-9999</td>
<td>374</td>
<td>08-1950</td>
</tr>
<tr>
<td>99/99/9999</td>
<td>346</td>
<td>04/08/1963</td>
</tr>
<tr>
<td>AAA 9999</td>
<td>312</td>
<td>Mai 1938</td>
</tr>
<tr>
<td>AAAAAAA-AAAA 9999</td>
<td>90</td>
<td>Janviers-mars 1999</td>
</tr>
<tr>
<td>99-99 9999</td>
<td>84</td>
<td>01-02 1993</td>
</tr>
</tbody>
</table>

Table 5.2: Ten most recurrent patterns for the date of collecting field

The same conclusion can be drawn from the results of the pattern analysis when applied on the dimension field. Measures are not standardized (both mm and cm are used) and apparently no rules were laid down regarding the syntax. Table 5.3 gives examples of the different ways of encoding the dimensions. As in the case of the problem with the dates, this incoherence renders the search process very difficult, not to say completely impossible. The output of this type of analysis can be used to
develop scripts for normalization and to build up value vocabularies.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Number of occurrences</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>99 A 99 AA</td>
<td>1,190</td>
<td>13 x 18 cm</td>
</tr>
<tr>
<td>999 AA</td>
<td>388</td>
<td>920 mm</td>
</tr>
<tr>
<td>999 A 999</td>
<td>382</td>
<td>573 x 100</td>
</tr>
<tr>
<td>99 AA A 99AA</td>
<td>196</td>
<td>37 mm x 16 mm</td>
</tr>
<tr>
<td>99 AA A 99AA A 99</td>
<td>107</td>
<td>52 cm x 25 cm</td>
</tr>
<tr>
<td>AA</td>
<td>14</td>
<td>72</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Examples of different patterns to describe dimensions

When applied upon the field “creditline”, which contains descriptors from a thesaurus, one can remark that each term is separated with one or more commas from the other terms. There exists an overlap with the content of this field with the title field.

5.4.3.3 **Histogram analysis**

The histogram analysis produces a histogram of the different values that exist for a specific metadata field. This analysis can be applied to quite a range of fields. Table 5.4 represents for example the titles that appear more than a thousand times throughout the collection. These data can serve as an excellent guide for discussions regarding the precision of the terms used in fields.

<table>
<thead>
<tr>
<th>Title</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>(empty)</td>
<td>5623</td>
</tr>
<tr>
<td>statuette</td>
<td>2043</td>
</tr>
<tr>
<td>panier</td>
<td>1800</td>
</tr>
<tr>
<td>bracelet</td>
<td>1792</td>
</tr>
<tr>
<td>collier</td>
<td>1376</td>
</tr>
<tr>
<td>masque</td>
<td>1324</td>
</tr>
<tr>
<td>groupe</td>
<td>1250</td>
</tr>
<tr>
<td>couteau</td>
<td>1073</td>
</tr>
<tr>
<td>sifflet en bois</td>
<td>1012</td>
</tr>
</tbody>
</table>

Table 5.4: Most frequent titles

As the histogram analysis gives a quick overview of how many times the same value appears throughout a column, it can also be easily used to
detect doubles. When applied upon the “objectid”, 348 doubles are detected, which represent approximately 5% of the total number of metadata records. In these cases, not only the content of the “objectid” field is identical, but also the content of the other fields.

By browsing through the results of the histogram analysis, a variation in the depth of indexation can also be detected. When applied on the “medium” field, we discover that the value “bois” is used 7,822 times, but along with a large number of unique values which are more specific, as for example “bois (vitex congolensis)”. From what does this difference in the depth of indexation depend?

Finally, the histogram analysis also sometimes allows to discover anomalies in an unpredictable manner. When applied to the field “objectcount”, the histogram analysis shows for example that 39 fields have the value “0”, which is a violation of domain range integrity, since an object must at least consist of one item.

5.4.3.4 Case analysis

The case analysis gives an overview of the use of capitalized and non-capitalized alphabetic characters. The application of this analysis is rather limited, but it still allows one to check the level of consistency of the metadata input.

Table 5.5 represents the results of the analysis when applied upon the field “region_old”, which shows that there are no general rules which are applied for the use of upper- and lower cases. As such, this has no effect upon the retrieval of information, but it illustrates the lack of documentation and coherence of the metadata creation process.

<table>
<thead>
<tr>
<th>Case type</th>
<th>Number of occurrences</th>
<th>Frequency (on the total number of non-empty fields)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed case</td>
<td>21,186</td>
<td>54.7%</td>
</tr>
<tr>
<td>All uppercase</td>
<td>14,889</td>
<td>38.5%</td>
</tr>
<tr>
<td>All lowercase</td>
<td>2,645</td>
<td>6.80%</td>
</tr>
</tbody>
</table>

Table 5.5: Use of upper- and lowercase characters

5.4.3.5 Length analysis

The length analysis calculates the number of characters used in a field. Again, this is a very basic query that is performed on the metadata. But its application definitely leads to interesting and unexpected results, as the
analysis allows to identify outliers and to match the content of the column with its definition.

When applied to the field objectnumber for example, the profiler informs that 69,718 values consist of 42 characters, and one value consists of 55 characters, although one visually sees by looking at examples of objectnumbers that the format of this field varies and never takes up 42 characters. The most frequent pattern “AA.9999.99.99” only consists of 13 characters, so where do these values come from?

Figure 5.9 shows the reason behind these values. A copy/paste of the data within a text editor such as Word reveals the formatting of the characters, and explicitly shows the whitespaces that are included within each value. The same phenomenon appears for the field date of production. Although the waste of memory within the database perhaps no longer is a big issue, the discrepancy between how the values are perceived and their true composition can pose problems for both the short and the long-term preservation of the metadata.

The use of the length analysis upon the “title” field also led to the discovery of incoherences in this field. The length of the titles varies from 2 to 142 characters, with an average of around 10 characters. This average is normal, but the so called “outliers”, which are the titles that consist of 100 or more characters, demonstrate again that there are insufficient rules that guide the metadata creation process. The same applies for the “description” field, where there are some extreme cases that contain 1999
5.5 Conclusions

The different analyses illustrated above clearly prove that simple and inexpensive data-profiling techniques can bring a great deal of problems or particularities within large sets of metadata to the surface easily. Table 5.6 summarizes the principal problems that were discovered for each field and propose actions that should be undertaken to correct the problems and prevent them from occurring in the future.

The high percentage of empty fields is the most striking outcome of the performed analyses. This end result still reflects the context of the initial creation of the first database, which was merely considered a practical instrument to follow the objects before and after the moving of the collection. With the purchase of the TMS collection registration software, an operation was started to correct an complete the existing metadata records. As the results from the analyses show, this operation is still far from completed. As the museum could only afford to buy a small number of licenses of the TMS software, the creation and management of the metadata cannot be distributed over a large group of employees. This financial bottleneck is one of the reasons why the department is currently migrating from the TMS software to CollectiveAccess, an open source collection registration software. However, the problem of empty fields is a recurrent phenomenon in historical databases.

The contrary phenomenon of the empty fields is also present, in the sense that certain fields contain metadata that should be split up over two or more fields. Jack Olson refers to this problem as field overloading, which is a common problem within legacy systems [110]. This problem occurs when descriptive practices change throughout time and needs rise to encode new types of metadata. As most commercial collection management systems are not easily customizable (see Section 3.4.1), catalogers often encode different values in the same metadata field, instead of adapting the metadata structure. The application of the pattern analysis upon the field “date of production” resulted for example in the discovery of metadata values such as “04.09.1992 (XVDS LST 2 : 122)”. In this example not only the date itself, “04.09.1992”, but also another type of metadata, “(XVDS LST 2 : 122)” is included. Especially when this occurs in a non-systematic way, it can become very difficult to split up the content over two or more separate fields in an automated manner.

What often worsens the situation is the creation of a structure within such a field with tabulations or spaces. These structures are only visible to the

---

7 Section 7.3.2 will present this software more in detail.
8 Boydens refers to the work of Peter Denley who has also worked on the problem of empty database records [25, p. 499].
### 5.5. CONCLUSIONS

<table>
<thead>
<tr>
<th>Metadata field</th>
<th>Detected problems</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objectid</td>
<td>5% of doubles</td>
<td>Elimination of the doubles</td>
</tr>
<tr>
<td>Objectnumber</td>
<td>Less than 10% does not correspond to the fixed syntax</td>
<td>Standardization of the minority of errors and creation of an input mask</td>
</tr>
<tr>
<td>Objectcount</td>
<td>39 cases with the value “0”</td>
<td>Manually check the records in question and build in an integrity constraint which prevents the encoding of the “0” value</td>
</tr>
<tr>
<td>Date_of_collecting</td>
<td>More than 80% is empty, incoherent syntax</td>
<td>Standardization and creation of an input mask</td>
</tr>
<tr>
<td>Date_of_entry</td>
<td>More than 50% is empty, incoherent syntax</td>
<td>Standardization and creation of an input mask</td>
</tr>
<tr>
<td>Date_of_production</td>
<td>More than 90% is empty, incoherent syntax</td>
<td>Standardization and creation of an input mask</td>
</tr>
<tr>
<td>Title</td>
<td>Small number of titles with a disproportional length</td>
<td>Build in a character limitation</td>
</tr>
<tr>
<td>Medium</td>
<td>More than 60% is empty, differences in the depth of indexation</td>
<td>Draw up documentation with indexation rules</td>
</tr>
<tr>
<td>Dimensions</td>
<td>More than 90% is empty, incoherent syntax</td>
<td>Standardization and creation of an input mask</td>
</tr>
<tr>
<td>Creditline</td>
<td>Almost 90% is empty, redundancy with the title field</td>
<td>Specify uniqueness of this field in regard to the title</td>
</tr>
<tr>
<td>Description</td>
<td>More than 90% is empty, elimination of the doubles</td>
<td>Draw up documentation with creation guidelines</td>
</tr>
<tr>
<td>Region_old</td>
<td>More than 40% is empty, inconsistency in case use</td>
<td>Standardization</td>
</tr>
<tr>
<td>Region_new</td>
<td>More than 40% Overlap with previous field</td>
<td>Study the relevance of the field and eventually delete it</td>
</tr>
</tbody>
</table>

Table 5.6: Overview of the problems and suggested actions to undertake upon the metadata fields

Human eye and cannot be used to extract the structure from the metadata automatically. This problem could be detected in the field “dimensions”, with for example values such as “1: L.: 755 mm B.: 110 mm. 2: L.: 530 B.: 230 mm.”. A human can immediately see that this field describes the length and height of two separate objects. During retrieval, these metadata are considered as one long unstructured row of characters within a
free-text field.

Apart from the problems regarding empty and overloaded metadata fields, there are clearly also problems with the consistency of the metadata input. Based upon conclusions drawn from the “pattern” and “length” analyses, both the syntax and the length of metadata varies in a disturbing manner in different metadata fields. The three fields that should represent dates are the most problematic. A standardized way of representing dates should be implemented within the database, which is the only option to allow efficient retrieval of records based upon this field. The length of values in the fields “title” and “description” also tends to differ largely. Titles and descriptions consisting of one or some words vary with values that consist of several paragraphs.

These results demonstrate how data-profiling can be used to obtain a general understanding of the formal quality of metadata and the detection of problem areas. The better understanding of the problems and incoherences within metadata records can guide future modifications and developments of the metadata scheme used by collection holders. This illustrates how the short time can have an impact upon the intermediate time. Chapter 6 and 7 will now build further and complement the results from this chapter by analyzing and demonstrating how cultural heritage institutions can act upon metadata quality problems.
Chapter 6

Correcting metadata with user comments

Contents

6.1 Introduction ........................................... 142
6.2 Possible strategies toward metadata correction ... 142
6.3 Cost-constrained metadata quality enhancement . 144
  6.3.1 Background of user-generated metadata ....... 145
  6.3.2 Case study: the image database of the National
        Archives of the Netherlands ..................... 148
  6.3.3 Case study: the Ross Archive of African Images . 156
  6.4 Conclusions ........................................... 163

Summary: User-generated metadata are gradually becoming a standard
feature of cultural heritage websites and databases. This chapter analyzes
which role user comments can play to augment and correct existing meta-
data. Two maximum variation case studies, which differ on the level of
expertise of the user comments, deliver empirical results that point out the
possibilities offered by user comments as a pragmatic method to enhance
metadata quality.
CHAPTER 6. CORRECTING METADATA WITH USER COMMENTS

6.1 Introduction

The methods to improve metadata quality differ largely from each application domain. The techniques of “data cleansing” that are applied within statistical databases, replace or remove data that are considered to disturb the overall informational value of the database [121]. The unique value of a single record is in this case not as relevant as the overall values that can be derived from the entire database. Such an error-tolerance on the individual level of a record is for example not tolerated within the administrative domain. In Section 2.3, I referred to the example cited by Isabelle Boydens, where a team of 300 people manually correct records from the social security databases. Within this application domain, every detected error has to be corrected, as the financial and social stakes are too high. This chapter will analyze how metadata are corrected within the cultural heritage sector and propose an innovative strategy of employing user comments to correct metadata. Two case studies will evaluate in which way user comments can correct published metadata of cultural heritage institutions and how relevant their content is toward other users.

6.2 Possible strategies toward metadata correction

The most straightforward approach toward metadata correction consists of manually checking records and updating them. This approach can be illustrated with the correction of metadata from the catalog of the University Library of Ghent. As already mentioned in 4.3.3, this library is busy with the retro-cataloging of their card-based catalog. The historic catalog was scanned in 2000, and with the help of OCR techniques the typed text from the cards was converted to digital text. This procedure however is not flawless, and errors can occur. Several members of the cataloging team are currently manually verifying the content of the scanned catalog card with the output of the OCR, correct the metadata where needed and update certain elements that are differently cataloged now then in the card catalog. Apart from being very resource-intensive, this kind of repetitive work is not always very rewarding. To motivate the cataloging team, a webpage (as demonstrated in Figure 6.1) has been created which uses Google Analytics to publish who of the employees has corrected the most records.

However, this approach of manually correcting metadata by professionals is mostly not a feasible option, as this operation is too costly. An interesting alternative in the same context of correcting OCR files is given by the reCAPTCHA tool. CAPTCHA, which stands for Completely Automated Public Turing test to tell Computers and Humans Apart, is one of the most
6.2. POSSIBLE STRATEGIES TOWARD METADATA CORRECTION

successful methods to prevent spamming scripts to abuse mail forms on the Internet to send spam messages. Before the form can be send, the sender has to identify a row of alphabetic characters that are encrypted within an image file, as illustrated in Figure 6.2. The text is easily readable by a human, but not for a script. The reCAPTCHA tool now uses image fragments of scanned text that cannot be converted to grammatically correct words. Thus, all of the text that is decrypted by Internet users contributes to the correction of historical digitized books.

A small number of research projects is also investigating the possibilities of automated metadata correction. The research team of the Language and Information Science department from the University of Tilburg has published some interesting papers on data-driven error detection and correction in textual databases from the cultural heritage domain [32]. A so-called horizontal correction algorithm tries to locate and correct inconsistent values within a database records, whereas a vertical correction con-

---

1For more information regarding the project consult http://recaptcha.net/.
CHAPTER 6. CORRECTING METADATA WITH USER COMMENTS

centrates on values that were entered in the wrong column. Supervised machine learning, based upon a controlled training set, replaces values that are identified as being incorrect with values that are more consistent with the values within the database record (see Figure 6.3). This approach definitively is interesting, but the success of the prototype as it is demonstrated within the research paper is largely to be attributed to the nature of their test collection, which consists of historical biodiversity metadata, with fields such as class, order, family, genus, species and subspecies. The same tool could not be applied upon other types of metadata which are less hierarchical or structured.

6.3 Cost-constrained metadata quality enhancement

Confronted with the sparse resources for the correction of metadata by professionals and the limited possibilities offered by automated means, I decided to investigate the possibilities offered by user comments. The changing role of the user, that has gradually shifted from a passive consumer of information toward a pro-active user that reorganises and manipulates data, has an increasing impact on traditional information retrieval.

A multitude of practical and methodological questions arise as popular web applications such as blogs, RSS and social bookmarking tools allow users to create and share metadata about online resources. More in particular, image databases from the cultural heritage sector increasingly offer users the possibility to annotate and comment on images. After a short
introduction to the issue of user-generated metadata, two maximum variation case studies will be presented which analyze the possibilities of user comments for the correction of metadata.

6.3.1 Background of user-generated metadata

As Eisenstein mentions in her study regarding the impact of the printed media, user-feedback already existed long before computer-mediated communications: “Sixteenth century publishers and editors created vast networks of correspondents, solicited criticism of each edition, sometimes publicly promising to mention the names of readers who sent in new information or who spotted the errors which would be weeded out” [47]. But generally speaking, the use of ICT and more in particular internet, has offered unprecedented possibilities for an ubiquitous interaction between a user and a resource.

A complete and coherent definition of what user-generated metadata precisely are remains tedious. User tags, as discussed in Section 3.5.2, and comments obviously enter the definition, but in a wider context, the log...
files of users who consult websites and databases also offer supplementary user-generated metadata on resources. One could compare these log files with real-life usage signs, which can offer very important information. Black rubber traces in the middle of the bend of a road can warn other drivers to slow down.

In the physical realm, documents and objects receive all along their life-span a certain patina, which the 19th century heritage crusader John Ruskin described as the “golden stain of time”, that offers indirect information on how a resource is being used [149]. In a sense, these traces can also be considered as user-generated metadata. What happens if we transpose this phenomenon to a digital online environment? Maes and Wexelblat specifically pose this question in their research paper “Footprints: history-rich tools for information foraging” and are experimenting with the creation of software tools for annotation, maps and paths that should enhance the navigation experience of users through large amounts of information [157].

Apart from these indirect, and mostly unconsciously, created metadata, I am here specifically interested in the users who deliberately create metadata. Folksonomies are probably the most popular example. They have been widely adopted by technology-aware individuals as a new method for organizing and retrieving online content. As mentioned in Section 3.5.2, some early adaptors within the museum sector are currently implementing folksonomies within their collection registration database. The intuitive and informal search options as offered by tagclouds can offer users with no specific search need a new browsing search interface.

Folksonomies can in no sense replace traditional cataloging and indexing, which are based on adequate vocabularies and performed by trained information professionals. Problems related to polysemy, synonymy and basic level variation result in an increased number of false positives and false negatives [64]. More research should be conducted regarding this issue but generally speaking, tags offer a very poor semantic value when used for describing images.²

I propose to focus the analysis of user-generated metadata by reviewing user comment. As comments are not restrained to a chain of one-word descriptors, they can offer a higher semantic value than tags. Secondly, large number of cultural heritage databases are currently offering this functionality, but no thorough analysis of user comments has been published.

²A good illustration of the poor semantic value of tags is demonstrated by the ESP-game (http://www.espgame.org/) which ames at “labeling all the images on the web”. It is a so called two-player game, where each player is randomly paired with another player and confronted with the same image. The goal is to guess what tags the other player is assigning to the image, in order to find a common ground tag. Most successful tags are very generic descriptors (“man”, “tree”, “red”) that offer few opportunities for specific image retrieval.
Throughout history, users have annotated and commented on resources. Let us think for example about the transcribers of manuscripts who frequently added glosses to the original work they were transcribing. As Verschaffel notes, historiographs in the 17th and 18th century also relied on the feedback of their readers to correct their work [154]. The rise of online databases has completely revolutionized the commenting and annotation process. However, insufficient research has been undertaken to examine the possibilities offered by these comments for information retrieval. Howard presented an interesting case study on the use of stories told by users to improve the description of historical images [78]. The main focus of the article is the documentation of an authoring tool which manages the user-generated comment. The Taskforce Archives\(^3\) from the Netherlands has published a book which presents an overview of the different web2.0 applications available for archives [39]. But again, this book does not contain any actual analysis of user-generated metadata.

One of the most interesting publications regarding user comments is the booklet the library of the University of North Carolina at Chapel Hill published on “Documenting the American South” [74]. This work contains the most interesting and striking comments the library received on this project. Excerpts from user comments are presented in five main categories: readers from the general public (general interest readers, readers with special needs, readers doing creative work, helping the younger generation), college and university readers (students, faculty), K-12 readers (students, teachers and school librarians), international readers. Apart from this typology, no other analysis is proposed.

As there is a clear lack of empirical research on the use of user comments within cultural heritage databases, I propose to study this phenomenon by analyzing two case studies. Within the debate on the utility of user-generated metadata, one of the central recurring elements is the opposition between the expert and the layman [82]. Therefore, I decided to perform analyses upon two maximum variation case studies, using exactly the same methodology. The first case study examines samples from user comments created by an undefined large public which consists mainly of laymen in the context of the image database of the National Archives of the Netherlands. This case study is then confronted with an analysis of user comments created by a very restricted number of world experts within the Ross Archive of African Images.

\(^3\)See http://www.taskforce-archieven.nl/
6.3.2 Case study: the image database of the National Archives of the Netherlands

The image database of the National Archives of the Netherlands was launched in 2004 and contains approximately 500,000 images. The collection of the former press agency Anefo forms the backbone of the database. Practically each Dutch news item from domains such as politics, sports, culture, economy but also the daily life activities from 1945 until 1989 can be illustrated with photographs of this collection. In this respect, the database represents a huge source for the contemporary history of the Netherlands. Most of the images made accessible through the database were digitized in 1997, and no longer correspond to current quality standards. The National Archives are now gradually replacing the older, low resolution files with high quality scans, which offer users the opportunity to zoom into the details of the images and to order high-quality reproductions.

From the detail view on the website, which represents the selected image accompanied with metadata such as description, date, keywords, collection, photographer, press agency and catalog number, users can click on a "comment" link which directs them to a form where free-text comments can be inserted. The user has to fill in his/her name and e-mail address and indicate whether this information can be published along with the comment on the website. The comment itself may not exceed 1,000 characters, but no other guidelines or restrictions regarding the content or style of the comment are given. A link on the homepage enables users to consult all the comments that have been published on the site. When a comment has been approved by a reviewer, it is published underneath the existing metadata.

As I just mentioned, the comments are reviewed before publication on the website. A specific employee of the collection management department judges individually the relevance of each comment within the back-end database. A comment can receive a different status, ranging from total deletion (the comment is seen as not pertinent and is never published on the website), provisional (newly arrived, not published), approved/to be treated (the comment is published on the website, but the existing metadata have not yet been updated), approved/treated (the existing metadata have been improved with the users comment, but the comment itself is no longer published on the website) and approved/published (the comment has been treated and remains published on the website).

These categories have been implemented to organize the treatment of the comments, but do not reflect a consistent analysis of the relevance of the comments. Therefore I decided to perform the analysis on the raw data as they entered the back-end database.

\[4\text{http://beeldbank.nationaalarchief.nl.}\]
6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

My goal is to evaluate the quality of the user comments. Despite the specific and very concrete nature of the comments, it is not feasible to investigate the truthfulness of the user comments as the objective correctness of the comments can not be analyzed in a deterministic manner. Following the “fitness for purpose” information quality definition presented in Section 1.1.3.2, my quality measurement will be the relevance of the comments for the user community. As noted by Svenonius, the principle of user convenience stands central within Anglo-American cataloging literature [137]. Foulonneau and Riley also point out the possibility of using search and browse behavior for managing recommendation systems in digital libraries [56].

Firstly, I need to determine the needs of the user community of the image database. The National Archives have not conducted any studies regarding the specific public they serve with the image database, nor on the specific user requests. Supplementary research to define the user needs issues is therefore necessary, for which I will rely on query analysis. This strategy has been widely adopted within the image research domain as an effective method to define user needs [34]. The mapping of the analysis of the user queries with the analysis of the user comments will give me an idea of the overlap and hence the pertinence of the comments toward user needs.

In order to structure the mapping between user comments and needs, I have used the faceted classification of Shatford (see Table 6.1) as a framework for the categorization of both the queries and the comments [129]. The process of categorizing heterogeneous content from user queries and comments into a limited set of abstract categories always implies some degree of violation of reality, but considerable preliminary testing with existing and new developed classification schemes proved the Shatford classification to be adequate for my purpose, as it focuses both on the level of specificity and the topics. Moreover, as this classification scheme already has been used in previous studies on the subject of user queries in image collections, it allows me to compare the results with previous research [5, 34].

6.3.2.1 Mapping user queries to the Shatford classification

Several search options are offered on the website. Visitors can either submit their query using the simple search field on the homepage, or they can use the advanced search mode which allows them to search simultaneously within specific fields. My analysis was based on the queries using the simple search field, as this is the most popular search option. During the period 05/04/2004 (date when the site went online) - 01/03/2006 (date when export was made) 465,124 searches were launched. Using a confidence interval of 5% and a confidence level of 95%, I obtained a
sample population of 384 queries, which I analyzed and mapped onto the Shatford classification.

This analysis resulted in the following results: S1: 17.50%, S2: 5.5%, S3: 57%, S4: 2.5%, G1: 9%, G2: 8.5% (G3, G4, A1, A2, A3 and A4 are not represented and some of the user queries could not be attributed to a category). The majority of users want to retrieve images related to a specific geographical location. Secondly, searches regarding specific individuals, groups or objects are also very popular. On the other hand there is a total lack of use of abstract query terms.

The two studies of P.G.B. Enser [5] regarding image requests in large non-domain-specific image databases confirm these outcomes. The larger part of queries refer to specific instances and unique items as object names and geographical locations, whereas more general and abstract concepts are not included. Studies focusing on requests within newspaper image archives reaffirm these results.

### 6.3.2.2 Mapping user comments to the Shatford classification

For the analysis I rely on an export from the administrative back-end database that stores all the metadata concerning the image database. The file contains 4,647 comments, sent in by users between the 5th of April 2004 and the 1st of March 2006. Each comment is accompanied by the name of the person that sent it in, the date of receipt and the status of the comment within the database, as described in the previous paragraph.

Using a confidence interval of 5 percent and a confidence level of 95 per-

---

<table>
<thead>
<tr>
<th>Iconography (Specifics)</th>
<th>Pre-iconography (Generics)</th>
<th>Iconology (Abstracts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who?</td>
<td>individually named person, group, thing (S1)</td>
<td>kind of person or thing (G1)</td>
</tr>
<tr>
<td></td>
<td>individually named event, action (S2)</td>
<td>kind of event, action, condition (G2)</td>
</tr>
<tr>
<td>Where?</td>
<td>individually named geographical location (S3)</td>
<td>kind of place: geographical, architectural (G3)</td>
</tr>
<tr>
<td>When?</td>
<td>linear time: date or period (S4)</td>
<td>cyclical time: season, time of day (G4)</td>
</tr>
</tbody>
</table>

Table 6.1: Shatford faceted classification
6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

As some users sent the same type of comment regarding a series of photographs after each other and the population of comments I worked on was sorted chronologically, I decided to perform a systematic sampling method, which ensured an even spreading of the sample over the population. As the comments are quite lengthy, they can easily incorporate more than one category of the Shatford classification, e.g., a comment can contain information regarding an individually named person, group, thing (S1) and individually named geographical location (S3). The analysis gave the following results: S1: 67.61%, S2: 18.87%, S3: 30.70%, S4: 20.56%, G1: 6.29%, G2: 1.71%, G3: 0.57%, G4: 0.29%, A2: 2.86% (A1, A3 and A4 are not represented).

6.3.2.3 Mapping of user queries with user comment

Figure 6.4 below demonstrates a general correlation between the content of the queries and the comments. Both queries and comments are highly motivated by interests in specific terms, use few generic terms and hardly any or no abstract notions. However, differences in popularity among the subcategories of specific, generic and abstract exist. The most popular queries concern individually named geographical locations (S3), whereas the most prevailing comments relate to individually named persons, groups or objects (S1).

This can be explained by the interest of users to find images of their hometown, street or even house, whereas users will most likely comment upon specific persons, groups or objects they know. In spite of these and other differences, I can conclude that the mapping of queries and comments demonstrated similarities in their content. The comments therefore help to fulfill a concrete information need from the user community.

6.3.2.4 Typology of the comments

Apart from the mapping of the comments and queries onto the Shatford classification, I also decided to deduct the most recurrent characteristics of the user comment. The same sample population of comments was used as with the Shatford mapping. This typology gradually emerged from the sample population during the manual analysis of the comments. This iterative process finally resulted in the following categories:

- correction of the displayed metadata (regarding spelling, identification of persons, event/action and geographical and temporal location): 45.58%

---

5Theoretically, the confidence interval is too high to be applied to small categories such as G2, G3, G4 and A2, but practical constraints did not permit a larger sample size [44]. Future research on this topic will be based using adequate sample sizes regarding all categories.
CHAPTER 6. CORRECTING METADATA WITH USER COMMENTS

- including narrative elements in relation to the image: 31.09%
- sharing the user’s personal history regarding the image: 8.95%
- mentioning a false or inadequate display of the image: 3.14%
- stating an opinion or judgement: 2.86%
- engaging in a dialog with the institution or other users, under the form of a question: 1.15%

These categories are neither exclusive or inclusive, meaning that a comment can belong to more than one category, but does not necessarily belong to one. Although the process of categorizing is inherently subjective, I do believe it allows a deeper understanding of the nature of user comments and their relevance toward the public.

Critical comments

The most important incentive for users to send comments is to express their disapproval with the published metadata on the website. 45.58% of all comments criticize the existing metadata, and propose a correction. Within this type of comment, 40.34% is relevant to an incorrect spelling of the metadata (mainly spelling of names of persons and locations), 10.92% to the identification of persons, 24.37% to the identification of an object, 15.97% to geographical location and 8.40% relevant to time. Figure 6.5 represents these figures in a graphical format.
6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

The National Archives recognize on the website that some image descriptions may be incomplete or contain errors, and therefore explicitly invite users to indicate potential errors and to send possible corrections. This new tendency of cultural heritage institutions to acknowledge and admit the sometimes low level of their published metadata was debated in Section 4.5.

I do not possess precise information regarding the validation of the corrections, but according to the National Archives, more than half of this type of comment is effectively used to correct the existing metadata.

Figure 6.5: Categories of user proposals to correct existing metadata

Including narrative elements in relation to the image

Another recurrent feature of the comments is their narrative character. More than 30% of the comments relate separate events in a narrative form or a story that enable a deeper understanding of the image. The sequence of metadata fields related to the identification of persons/objects depicted, location, date, etc that are published by the National Archives on the site does not imply an understanding of images. User comments can introduce some narrative aspects within the database by making connections between events, persons, locations and temporal specifications. Figure 6.6 gives an example of a user who sketches the very detailed historical background of a tram compartment used by the German army during the Second World War.

I can relate here to Lev Manovich, who noted the rise of the database as a cultural form and its tension with narratives:

As a cultural form, the database represents the world as a list of items, and it refuses to order this list. In contrast, a narrative creates a cause-and-effect trajectory of seemingly unordered items (events). Therefore, database and narratives are natural enemies. Competing for the same territory of human culture, each
claims an exclusive right to make meaning out of the world.[97, p 225]

Here it is not so much the content of the information, but its form that is of interest. The limits of the dissertation do not allow us to go deeper into the subject, but recently, a whole new research area emerged that studies the impact and possible benefits of narrativity and storytelling within the domain of information representation. For a more elaborate vision on the impact of narrativity within the representation of cultural heritage, the work of Wendy Duff can be consulted [45].

Figure 6.6: Example of a user comment that adds a socio-historical context to an image

Linking the user’s personal history to an image

A minority of comments consist of users that disclose their personal relation to the image content. Discovering themselves, family, friends, their old hometown or simply a familiar scene on historical photographs, users are enticed to communicate their personal experiences. Figure 6.7 illustrates the highly personal story of user who describes her impressions regarding the collecting of clothes and toys for victims of the water flood that struck the Netherlands in 1953.

In which sense are these personal comments meaningful/useful to other users? Further research has to be conducted in order to evaluate the long-term quality and pertinence of this kind of personal information.

Stating an opinion or judgement

Surprisingly few comments contained explicit opinions or judgments of users. Figure 6.8 gives an example of a user who takes the image of an old football field as a point of departure to complain about the actual presence of an unappealing block of houses where a football field used to be.
Again, we can ask ourselves what is the pertinence of metadata intrinsically related to an individual. The same phenomenon appears with social tagging where tags identifying qualities or characteristics (e.g. “free”), incorporating self reference (e.g. “my stuff”) and individual task organizing (e.g. “to read”) are used. More research should be conducted in order to know if these tags are effectively used by other users as an informal recommendation system. Analogies can be made with the use of user bookreviews on the website of the online book retailer Amazon.

Engagement in a dialog with the institution or other users

A fraction of the comments contain questions or invitations toward the institution or other users to obtain more information regarding a specific image. Figure 6.9 gives an example of such a question. Sometimes dialogues between users take place, which transforms the comments page into a forum where users can interact with one another. Even if the technology for implementing a forum within a website is widely available at a low cost, there are very few heritage institutes that offer such a feature [22].
CHAPTER 6. CORRECTING METADATA WITH USER COMMENTS

Figure 6.9: Example of a user asking a question

Mentioning a false or inadequate display of the image

A last category of comments consists of users mentioning an incorrect display of images. A recurrent problem with high-volume scanning projects of photo negatives is the appearance of mirrored images, which are very hard to detect. When no text is displayed on the image, one has to personally know the scene or object depicted in order to detect the mirrored view.

6.3.3 Case study: the Ross Archive of African Images

The Ross Archive of African Images is a database of over 6,000 illustrations of figurative African objects published between 1800 and 1920 in books, periodicals, catalogues, newspapers, and other publications. Almost all of the material comes from the James J. Ross library, only a small number of illustrations has been scanned from copies held in other libraries. The chronological limitation of the content of the database to the period between 1899 and 1920 is guided by the idea that from 1920 on, the first falsifications appeared within the world of African art. This chronological limitation tries to offer a guarantee of the authenticity of the objects which are represented within the database.

In contrast with the image database of the National Archives of the Netherlands, the resources and the metadata of the Ross Archive of African Images are highly specific. The access to the database is also limited to a very restricted number of worldwide experts on African Art. All of the original texts that accompany the illustrations are included within the database in their original language, but a translation in English is always provided. Figure 6.10 represents a typical record from the database and illustrates the highly specific content of the metadata and of the comments sent by the users.

The manager of the database has provided us with an export of 5,401 comments that have been sent to the website. Using a confidence interval
6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

of 5 percent and a confidence level of 95 percent, a sample population of 359 comments was obtained and manually analyzed.

6.3.3.1 Mapping user queries to the Shatford classification

The website of the Ross Archive of African Images offers both a simple and an advanced search. The latter offers researchers the possibility to search specifically in thirteen different fields of the database, as for example the name of the illustrator or the techniques used for the illustration.

My analysis was based on a file which contained 4,046 queries, resulting from both the simple and the advanced search field, logged on my demand by the database administrator from 27th May until 21st July 2008. Using a confidence interval of 5 percent and a confidence level of 95 percent, I obtained a sample population of 351 queries, which the author analyzed and mapped onto the Shatford classification.

This analysis resulted in the following results: S1: 81.22%, S2: 1.5%, S3: 13.08%, S4: 0.5%, G1: 14.86% (G2, G3, G4, A1, A2, A3 and A4 are not represented and some of the queries could not be attributed to a

Figure 6.10: Example of a record from the Ross Archive of African Images
category).

An overwhelming majority of users wants to retrieve images related to an individually named object. This is related to the high presence of inventory numbers within the query log file. Otherwise, searches on specific geographical names and on generic kind of things, as for example “pottery”, also take up an important part of the user queries. As it was the case with the image database of the National Archives, a total lack of use of abstract query terms appears.

6.3.3.2 Mapping user comments to the Shatford classification

The same mapping of user comments to the Shatford classification has been applied to the comments send to the RAAI. The results are quite different: S1: 96.94%, S2: 3.62%, S3: 12.26% and S4: 3.06%. Again, one comment can belong to different categories. The other parts of the Shatford classification are not represented. These figures show an overwhelming importance of comments that relate to the specific content of the images or of the metadata, and more in particular to specific persons, groups or objects that are depicted. This is in clear relation with the high-level of knowledge that the researchers possess on the objects contained within the database. The photographs also only represent highly specific content, in contrast with the very broad scope of the image database of the National Archives.

6.3.3.3 Mapping of the user queries with user comments

In analogy with the first case study, I discovered a general correlation between the content of the queries and the comments. Figure 6.11 demonstrates this correlation. Both queries and comments are highly motivated by interests in specific terms, but use few to no generic terms and abstract notions.

These results are clearly affected by the nature of the users of the database, which are basically the same limited number of African art specialist that can send comment. These people are world-wide experts in the domain of African art and have a very clear idea on what they are searching in the database and how they can assess the information it beholds. The intensive use of inventory numbers is an example of the highly specific level of their searches. This implies that the researchers already have a very specific image in mind before they start searching.

In conclusion, the mapping of queries and comments demonstrated similarities in their content. The comments thus help to fulfill a concrete information need from the user community.
6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

6.3.3.4 Typology of the comments

Quite a different typology resulted from the analysis of the comments of the RAAI than from the National Archives. Again, this typology gradually developed when manually analyzing the sample of user comments. Only three distinct categories emerged from the analysis:

- complementing the existing metadata: 75.77%
- linking the object with other information resources (catalogs, publications or websites): 61.27%
- correcting the existing metadata or mentioning issues with the displayed images: 18.38%

An important overlap exists between the categories, in the sense that a large part of the user comments which give supplementary information or that question the truthfulness of the existing metadata are backed up by references. This uniform character of the user comments stands in close relation with the coherence of the user group of the database, which is built up of a small network of researchers.

Comments giving additional information

In analogy with the observations I already made when comparing the comments with the queries, I noticed that more than 75% of the comments complemented the existing metadata with information on the object (= the category “what”) that is depicted. As the database specifically contains images that illustrate African art objects, this observation is quite obvious.
The second largest category of complementary metadata gives additional information in relation to the geographical location of the object described in a record. Only a small minority of comments falls into the “who” and “when” category. Figure 6.12 represents graphically on which levels user comments complement metadata.

![Figure 6.12: Categories of user comments which complement existing metadata](image)

**Including references on the object from other information sources**

It was somehow surprising to observe such a high frequency (61.27%) of comments within this database that contained references to catalogs, inventories or publications. The act of using references to back up assertions made in comments is clearly in relation with the scholarly background of the contributors, who have the habit of using references. Figure 6.13 gives an example of the type of references used.

When assertions are made without any references to articles or other publications, the contributors are very careful not to give the impression that they are stating something that has been verified.

Most comments refer to scientific publications or official catalogs, but a considerable amount of users also refer to online information resources such as Ethnologue, an encyclopedic reference work that catalogs all of the world languages.6

**Adding corrections**

The fact that only 18.38% of the comments consist of actual corrections of the metadata is probably in relation to the scientific character of the

---

6.3. COST-CONSTRAINED METADATA QUALITY ENHANCEMENT

The most frequent type of corrections (65.15%) consist of comments declaring that some part of the identification and/or description of the object is incorrect. The second most frequent type of corrections oppose the geographical metadata (21.21%). The other types of corrections are only marginally present: correction of spelling (6.07%), the way an image is reproduced (4.55%) and the identification of persons or tribes (3.02%).

Figure 6.13: Use of references to back up the validity of user comments

Figure 6.14: Categories of user comments which correct existing metadata
Some researchers go into great detail to explain why they criticize the current interpretation of an object or document. The following researcher states nine different reasons why he questions the content of a certain sketch that pretends to represent a realistic scene:

I have no confidence in this sketch or in the similar sketch in Frobenius 1909. It seems to show the performance of the Baga Dimba mask. But there is nothing in Baga ritual, cosmology, masking, or dance that corresponds to anything in either of these sketches or in the accompanying text in the 1909 publication. The contradictions are many: 1) multiple Dimbas never danced together. Although each quartier in the village had its own Dimba mask, and therefore any village might have 3-5 Dimbas, each always appeared separately, at a separate occasion. They were competitive, not collaborative. 2) The 1909 sketch, especially, shows the dancers faces clearly just below the neck of the mask. But the Dimba mask has a lower half representing shoulders and female breasts, and the dancers face is completely hidden behind the breasts, where he sees out only through two small holes between the breasts. 3) Dimba never danced at any male initiation. 4) No male initiation has ever been described to me remotely matching the image or the description in the 1909 publication, in which the initiates are tied to a pole, on top of which is a human skull festooned with feathers. Granted, to learn the secrets of male initiation is rare and difficult, but I gathered quite a bit of information, enough to know what kinds of initiations existed. 5) The sketches show a large bonfire casting direct light on the initiates bodies and dark shadows behind them, indicating that this is night. But Dimba always danced in the bright sunlight of mid-day, and this was critical to her identity. 6) Dimba never danced in the bush, but always in the town. 7) The Dimba dancer never carried a switch, because the dancers hands needed to grip the two front legs of the very heavy mask. 8) I have never heard of any Baga ritual called Juju Nkali, and the construction of the word itself does not seem Baga. And 9) I have fairly exhaustively searched the archives on three continents for information on missionaries who worked in the Baga region, and I have never come across the name of Keil. So I think Frobenius was confused, and the sketches seem to be a fantasy.\footnote{This quote has been copied from the RAAI database.}

Only a minority of the comments point out errors within the metadata or the way the object is depicted on the images. Figure 6.15 illustrates the comment of a user who points out that the digital image represented within the database is the product of a merger of two analogue images.
6.4. CONCLUSIONS

This chapter focused on an evaluation of the potential of user comments for the correction and enhancement of metadata quality, with the help of two maximum variation case studies: the image database of the National Archives of the Netherlands and the Ross Archive of African Images. I decided to confront two case studies which are very different on one specific dimension: the level of expertise of the user comments. As the debate on the utility of user-generated metadata focuses heavily on the difference between experts and laymen, this specific characteristic was chosen to differentiate two case studies. The image database of the National Archives of the Netherlands invites everyone to submit comments, whereas only a very limited number of world experts are invited to submit their comments to the Ross Archive of African Images.

In both case studies an initial assessment was made of the adequacy of the user comments by confronting them with the content of the user queries. As I was confronted with empirical data whose correctness cannot be verified in a deterministic manner, a pragmatic approach was adopted consisting of a mapping of the content of the user comments with the content of user queries by using the Panofsk/|Shatford classification.

In the two case studies both queries and comments are highly motivated by interests in specific terms. In the case of the image database of the National Archives, there is also a small part of the comments and queries that focus on generic terms and a few on abstract terms. The total absence
of those two categories within the Rossbook Archive of African Images is in relation with the highly specific content and the level of expertise of its users. In both case studies a clear relation between the user queries and comments can be found, so therefore I concluded that the user comments are relevant toward the user community.

In a secondary phase, I investigated how many of the user comments effectively propose corrections and/or enhancements of the existing metadata, by developing for each case study a typology of the user comments. Here I did not rely on a pre-existing typology but developed in a bottom-up approach a new typology for each case study by manually analyzing samples of the comments.

The development of the typologies for each case study clearly showed that user comments can play an important role for adding extra metadata or for obtaining feedback on the validity of the existing metadata. Almost half of the comments from the large public and practically all of the comments from the experts fulfilled these roles. The second important outcome from the typologies was the low rate of personal and subjective comments from the general public.

Even if the analysis of the user comments from the image database of the National Archives of the Netherlands only showed a low percentage of highly personal and emotive comments, the question remains how cultural heritage institutions have to deal with this kind of interaction from their public. The combination of “official” metadata, meaning professionally created descriptions, and user-generated metadata under the form of comments or social tagging result in two different levels of indexing, in the sense that the official documentation developed by the institution is confronted with the sometimes subjective interpretations of users.

This issue joins other tendencies regarding the role of users within the debate on metadata quality. Based upon the analysis of the user comments, which result from an interaction between the users and the existing metadata, recommendations can be made which will influence the mentalities regarding the interactivity of users which is debated within the “temps long”. Section 8.1.4 from the conclusions will bring together the different issues regarding user interactivity mentioned throughout the dissertation and develop a final position on this question.
Chapter 7

Monitoring needs with search interfaces

Contents

7.1 Introduction ............................................. 166
7.2 Conceptual model of dynamic search interfaces ... 167
  7.2.1 Gathering information on user needs ......... 167
  7.2.2 Considering user actions as metadata of user needs .............................. 167
  7.2.3 Monitoring metadata quality instead of direct marketing ...................... 168
7.3 Case study: the September 11th Memorial and Museum ............................... 170
  7.3.1 Description of the September 11th Memorial and Museum .................. 171
  7.3.2 Description of CollectiveAccess .................. 173
  7.3.3 Implementation of the dynamic search interface prototype .................... 175
7.4 Conclusions ............................................. 178

Summary: This chapter provides the third and last strategy for the evaluation of metadata quality. Confronted with the importance but also the difficulties of defining user needs sketched in the first part of the dissertation, an original concept for automatically detecting user needs is developed. With the help of the developers of CollectiveAccess, a dynamic search interface is programmed and implemented within the collection registration database of the September 11th Memorial and Museum.
CHAPTER 7. MONITORING NEEDS WITH SEARCH INTERFACES

7.1 Introduction

The third strategy presented in this chapter offers the necessary complementary approach to finalize the metadata quality cycle offered in the operational second part of the dissertation. The work on data-profiling and user comments from Chapter 5 and 6 respectively demonstrated how to obtain a diagnosis of the formal quality of metadata and how user comments can be used to correct existing metadata. The third and last strategy now presents a prototype to monitor in a seamless way over time the adequacy of existing metadata fields within the environment of a collection registration database, by offering the users of the database the possibility to customize their search interface in an intuitive manner.

Just like the two other strategies, the dynamic interface has been conceived as an outcome of the confrontation between the theoretical first part and the practical constraints with which metadata practitioners are confronted. Following the “fitness for purpose” definition, the adequacy of the metadata toward its users stood central throughout the different chapters of the first part of the dissertation. Section 2.2 also pointed out to the difficulties of defining user needs within the cultural heritage sector: user needs are hard to obtain, evolve throughout time and partly even do not exist yet, as we cannot predict future uses.

Taking into account these difficulties, I decided to develop and implement in collaboration with the head of the development team of an open source collection registration software an innovative tool which allows users to add and delete the metadata fields of their choice within the search interface. This approach prevents the possible pitfalls regarding the definition of user needs which were discussed in Section 2.2, as the statistics obtained by this tool reflect the actual needs in a much less biased way than results obtained by questionnaires or interviews. Due to the automated and intuitive character of the tool, it requires no substantial resources, which is also an important advantage compared to other user need approaches.

The concrete implementation of the dynamic interface also underlines the power of the conceptual framework of the stratified time which was presented in the fist part of the dissertation. In contrast to the original framework developed by Fernand Braudel, where the interaction consists of a unidirectional impact of the longer sequences upon the shorter time levels, Isabelle Boydens has argued that in the context of information systems, the shorter time levels can also influence the longer sequences. This modification of the stratified time concept is demonstrated by the implementation of the dynamic search interface. The day-to-day interaction between metadata and their users is used to develop recommendations to modify the structure of the metadata, who are discussed within the intermediate time, and to guide the development of a long-term digitization policy.
7.2 Conceptual model of dynamic search interfaces

7.2.1 Gathering information on user needs

One of the most frequently used methods to obtain information from users and their needs are surveys. Web-based forms and email greatly improve the speed and lower the costs of organizing surveys, but their effectiveness is questionable. The experience of the MOAC project mentioned in Section 2.2.2 shows that only a limited set of data can be gathered. Direct interaction with users, either through face-to-face or telephone interviews, or through observation in a usability-testing environment, is also a recurring method. The preparation, set-up and analysis of interviews and observations are very resource intensive. Additionally, these methods only deliver results which represent a “snapshot” in time.

More automated means of capturing user needs are preferred within this dissertation. Once an automated data capture method is configured, it can be run at any time at no extra cost. This means that the outcomes of the analyses can be monitored over time, instead of relying on snapshots of a certain situation at a specific moment in time. This approach allows to discover trends and evolutions in user needs by monitoring the actual use of metadata fields over long periods of time. As Foulonneau and Riley state, “metadata quality control is an iterative process that never ends” [56].

7.2.2 Considering user actions as metadata of user needs

The success of popular web applications such as Facebook or Last.fm depends for a great extent on the intuitive manner in which these applications gather information on user needs. Both software tools rely heavily on the analyses of metadata on user needs and preferences to offer customized services. When installing Last.fm for example, the application asks if it can analyze the metadata of locally installed media players such as iTunes. These existing metadata on the musical taste of the user are then uploaded within the user profile on Last.fm so that a new user immediately has a profile which reflects his taste in music. Once the software is installed, all of the metadata of the music files played on the computer are logged automatically in the background by Last.fm and used to continuously refine the profile of the user.

The social networking site Facebook handles the same approach of unconsciously aggregating metadata from users. The first time a user logs

---

1Facebook and Last.fm are examples of social networking websites. Facebook (http://www.facebook.com) is centered around the use of images, whereas Last.fm (http://www.last.fm) is used to connect people through their taste in music.
in, the application asks if it can import all of the e-mail addresses from different types of accounts, and then checks who is already on Facebook so that the new user can send a mail in an automated manner to all of these contacts. Once a user is logged in, the actions undertaken upon the platform are used to render his profile more detailed, and are distributed to all of the people in his network.

These two examples demonstrate how metadata describing user needs and interests can be gathered in an automated manner, by relying on small actions performed by the users in an unconscious manner. Basically, the user never has to actively encode metadata in any way, but still manages to receive and manage content that specifically fits his needs. The success factor of these applications resides in how users can be persuaded to spend time customizing their profile, by clearly offering them an added-value service. Figure 7.1 illustrates what kind of services are built on top of the personal profile of a user. The top part of the screenshot displays the audio tracks that were played by the user, which are then used by the application to recommend other music, but also events such as concerts that take place around the home area of the user.

I decided to adopt a similar approach to gather information about how and which metadata fields are effectively used within a documentary database. As previous chapters have demonstrated, users are not inclined to participate in initiatives such as surveys or questionnaires where they have to take the time to actively enter information that is of no direct use to them. It is therefore crucial to offer a direct added-value to users if I want them to interact with an application which will allow me to obtain metadata regarding their needs.

Concretely, the idea behind the dynamic interface is that the user will be confronted with a default search interface which can be very intuitively configured to the taste and needs of the individual user. I inspired me here by the way blocks of information can be “dragged and dropped” on the iGoogle interface. This application (illustrated in Figure 7.2) gives the possibility to customize the different information blocks, as for example the headlines of favorite newspapers, stock exchange information or local weather reports, which appear on a homepage. In a similar way, the user of the dynamic search interface will be able to add and to delete metadata fields.

7.2.3 Monitoring metadata quality instead of direct marketing

The customization by users of their profile on applications such as Facebook offers them the direct advantage of having a more detailed profile page. The gain toward the managers of Facebook is tremendous. Never in
7.2. CONCEPTUAL MODEL OF DYNAMIC SEARCH INTERFACES

the history of marketing have potential customers given away so much information in a pro-active manner about their needs and preferences [161]. This explains why social networking applications are being sold for prices which seem exorbitant at first hand.

The interactive interface does not deliver information to be used for di-

Figure 7.1: Screenshot of the Last.fm application illustrating the services built on top of automatically inferred user needs
CHAPTER 7. MONITORING NEEDS WITH SEARCH INTERFACES

rect marketing purposes, but behind the direct interest towards users of having a customizable search interface lies also another motive. The real incentive behind the dynamic interface is the ability to gather information on which metadata fields are actually being used. The seamless and automated way of gathering this information will allow to monitor changes in user needs and perform statistical analyses over long periods of time.

7.3 Case study: the September 11th Memorial and Museum

Now that the idea behind the dynamic search interface is developed, the concept needs to be implemented within an actual search interface used to perform queries on the database of a cultural heritage institution. In order to obtain a maximum result out of the interaction between the prototype and its beta users, two presuppositions are formulated. Firstly, as the tool is still in a prototype phase, heritage institutions will not be very motivated to implemented the tool directly within their public front-end. On the other hand, the implementation within the back-end of a collection registration database offers the advantage to test the tool internally with the cataloging team. As catalogers work several hours a day with the collection registration database of an institution, they have a high interest.
The next step is to determine the type of cultural heritage institution where the prototype could be implemented. One of the basic ideas behind the dynamic search interface is to monitor changes in the day-to-day use of metadata within a cultural heritage institution. As new types of objects enter the collection of an institution, the catalogers need to adapt their descriptive practices and their search behavior. Newly founded institutions which are still in the midst of building up their collection would therefore provide an excellent test bed for the dynamic search prototype.

Based upon this set of criteria, I approached the September 11th Memorial and Museum with the question to implement the prototype within their collection registration software, which they kindly accepted. The following sections will present in detail the museum and the collection registration software they use.

7.3.1 Description of the September 11th Memorial and Museum

The September 11th Memorial and Museum in New York manages objects and digital assets related to the events of September 11th, 2001. The collection is largely still under construction and is growing at a fast pace. Figure 7.3 illustrates the homepage of the museum's website.

Several reasons justify the choice of this museum as a case study to illustrate the third operational strategy of the dissertation. Most importantly, the highly evolving and very recent character of the September 11th Memorial and Museum makes its collection a perfect case study to analyze the adequacy of the dynamic interface tool. New object types enter the collection frequently and oblige the collection managers to rethink the existing metadata scheme. When new object types as for example destroyed firetrucks or steel construction pillars of the World Trade Center enter the collection, the existing metadata scheme is not necessarily capable of describing these new types of collection holdings. The implementation of the prototype of the dynamic search interface can monitor how the search behavior evolves in this context by analyzing which specific metadata fields are actually used and which not.

Moreover, the September 11th Memorial and Museum very much illustrates the new nature of cultural heritage as it is understood in the 21st century. This case study demonstrates as no other the changes that were described in Section 1.1.3.4 where I analyzed the evolutions in the contemporary thinking on cultural heritage. What happened yesterday, can

---

2See the museum’s website on http://www.national911memorial.org/.
already claim its role within our collective heritage. The World Monument Fund placed in 2002 the area of the so-called “Ground Zero” on the list of endangered cultural heritage sites [60].

Apart from illustrating the new temporal character of cultural heritage, the museum also demonstrates the evolved social embedding of cultural heritage. The museum actively invites all of the individuals or communities concerned by the terrorist attacks to take part in the construction of a common heritage around the events of September 11th. This confirms the evolutions described in Section 2.3.1, where I gave an account of how cultural heritage institutions currently use the Internet to build virtual communities around their collections. The direct presence of merchandising and fundraising activities, illustrated by the “Contribute” link on the homepage represented in Figure 7.3, also demonstrates the evolutions described in Section 2.3.2. Cultural heritage institutions are increasingly obliged to place themselves as regular players on the marketplace to obtain supplementary incomes.

The existing collaboration between the museum and the CollectiveAccess project, described in the next section, evidently also facilitated the actual implementation of the dynamic search interface prototype within a short period of time.
7.3.2 Description of CollectiveAccess

In 2007 I established a collaboration with the development team of the CollectiveAccess project. CollectiveAccess is a general-purpose collections management system, intended for use with a wide variety of materials. Current users include representatives from many fields, including fine art, anthropology, film, oral history, local history, architecture, material culture, biodiversity conservation, libraries, corporate archives and digital asset management. The website of the project, illustrated in Figure 7.4, gives a detailed overview of the different user groups of the software.

![CollectiveAccess](image)

Figure 7.4: Homepage of the CollectiveAccess project website

The most important features concerning metadata management are:

- **Completely web-based user interface**: this implies that metadata input can be very easily distributed amongst a large group of catalogers or external experts.

- **Customizable metadata schemes**: in addition to the standard set of CollectiveAccess fields representing concepts applicable to anything that can be cataloged - like “accession number” - sets of attributes, functioning as repeatable custom fields, may be defined. These sets can map to established metadata standards such as Dublin Core, Darwin Core, VRA Core 3.0, CDWA Lite and others. Attribute sets may be type-specific: they can be defined such that they are only available...

---

3 See the project’s website on http://www.collectiveaccess.org.

4 These features are also described in [151].
for specific types of catalogued items (ex. photographs, video tapes, films). They may also be repeating, and it is possible to impose an intrinsic data type (text, integer or floating point number, date) as well as bounds and pattern-based input validation rules.

- **Automatic extraction of technical metadata**: technical metadata that reside within the file headers of media files are automatically recognized and stored within the database when uploaded media files.

- **Controlled vocabulary support**: extensive support for authority lists and controlled vocabularies. A tool is included to import the Getty Art and Architecture Thesaurus (AAT) data files.

Apart from the technical features of the software, one of the crucial incentives behind the collaborating with the CollectiveAccess team is the fact that the application is available under the GNU General Public License (GPL) version 2. The use of the GPL guarantees, as mentioned on the website, in perpetuity a user’s right to (a) use CollectiveAccess without charge, (b) redistribute CollectiveAccess to others without limitation and (c) freely modify CollectiveAccess to meet their needs or the needs of others.

As Section 3.4.1 has demonstrated, the closed character of proprietary collection registration software can have a negative impact upon metadata quality. As users have no access to the source code of the application, they are obliged to work with the vendor in order to perform any modifications of the software, even when the users of the software are unsatisfied with his services. On the other hand, it is a fallacy to believe that the implementation and the use of open source software is free. Most cultural heritage institutions do not have sufficient internal ICT resources to assure the customization of the software and to give training to the collection management staff. In the case of open source software, users can work with free lance consultants of their choice. The user community that exists around an open source software can also contribute to the documentation process and help out other users with specific questions.

After a detailed comparison of different commercial and open source solutions, the September 11th Memorial and Museum chose in April 2007 CollectiveAccess to manage their holdings. The existing collaboration between Whirl-i-Gig, the company which developed CollectiveAccess, and the museum tremendously facilitated the implementation of the dynamic search interface.

---

5 See the company’s website on http://whirl-i-gig.com.
7.3.3 Implementation of the dynamic search interface prototype

The general idea behind the dynamic search interface and its implementation within CollectiveAccess rose during the research I performed in collaboration with Seth Kaufman, director of Whirl-i-Gig and head of the development team of CollectiveAccess, on a paper and a presentation for the International Conference on Dublin Core and metadata applications, held in September 2008 [151]. During a meeting in November 2008 at the September 11th Memorial and Museum in New York, Seth Kaufman and I presented the concept and the implementation of the dynamic search interface to Adena Langer, cataloger of the museum and asked whether we could implement the future prototype within the museum’s database. Following the official approval of the museum in December, Seth Kaufman effectively programmed the dynamic search interface prototype and implemented the prototype within the collection registration database of the museum in December 2008 - January 2009.

Due to the highly sensitive and personal nature of certain collection items of the museum, no direct access to the database of the museum could be given to me for privacy reasons. In order to follow up the implementation process of the prototype, the database manager provided me with screenshots from the database and exports of the logfiles with some preliminary results from the interaction between the catalogers and the prototype. The prototype was also installed within a dummy version of a CollectiveAccess database for me to experiment with.

Figure 7.5 illustrates how the dynamic search tool is implemented within the interface of the collection registration software. On the right hand side of the search fields is a box, illustrated in detail by Figure 7.6, which contains a drop down list with the different available forms and the available form fields. New forms can be created and existing ones deleted. Users of the tool can define several search forms. Having different customized search forms can be an interesting feature for users which need to perform regular searches for different types of collection holdings. When a new form is created, fields have to be added by clicking on the field name in the right hand column. Once a field is on the search form, it can be removed by clicking on the tick-box next to the field name. When defining a new form, users can decide to give other users of the collection registration database access to the specific form by indicating that it is a system form.

The first time a user logs into the collection registration database, he has access to a default advanced search (illustrated in Figure 7.5) which includes the following fields:

- **object title**: free text field
- **full text search**: searches through every field
Figure 7.5: Illustration of how the dynamic search interface is implemented within the collection registration database

- **object has media**: only returns records which have media (images, audio, video, etc)
- **date**: free text field
- **object classification**: drop-down list describing the type of object (archival, moving image, sound, etc)
- **collection category**: drop-down list describing the type of collection (National September 11th Memorial and Museum, Port Authority Hangar 17 Collection, Storycorps Collection, etc)
- **object vocabulary terms**: free text field
7.3. CASE STUDY: THE SEPTEMBER 11TH MEMORIAL AND MUSEUM

Figure 7.6: Detailed view on how a new search form is defined

- **has flag**: drop-down list with user defined values (damaged, needs conservation, to be located, etc)
- **object status**: drop-down list describing the status of the record ("new", "first edit", "completed", etc.)
- **object modified by user**: drop-down list with the names of the database users which have modified records
- **object modified on date**: free text field
- **storage location**: free text field

The dynamic search interface now offers the opportunity for a user to configure one or more search forms which only contain the search fields of his interest. Figure 7.7 illustrates a search interface that could be build by a user who has to perform regular searches for video on different aspects related to the 9/11 events, which is ready to be published and distributed online. The field “classification” allows to select only records which have video as an object, the “status” field gives the opportunity to only select records which are ready to be published on the website. The field “object title + content description” finally allows to select the content of the video by searching on keywords figuring in the title or object description of the record.

The different actions undertaken by the users of the dynamic search interface are logged by the system. The following information is stored when users create or modify a form with the prototype:
Figure 7.7: Example of a customized search form

- Date and time of event
- The form
- The user
  - The type of event: form created, form deleted, field added to form, field removed from form
  - What the field was

As the prototype has just been developed and implemented at the museum in December 2008 - January 2009, no significant log files are already available to perform statistical analysis upon. I will monitor the log files throughout the next months and plan to publish an article on the trends that will emerge from the log files. Within the log files, particular attention will be given to which fields are added and which ones deleted, and how these operations evolve throughout time.

7.4 Conclusions

Chapter 7 presented the third original strategy towards the correction of metadata quality within cultural heritage institutions. Even more than with the two previous strategies, the observations and the outcomes of the first part of the dissertation logically led to the development of the approach presented in this chapter. The importance but at the same time the difficulties of measuring user needs which were discussed in Section 2.2 led me to the idea of developing a methodology and tool to monitor user needs regarding metadata in an automated manner.

The success of popular social web applications such as Facebook or Last.fm inspired me to transpose their way of aggregating information on user needs by analyzing the actions of users. Facebook and Last.fm demonstrate how users can be convinced to spend time customizing their user-profile by offering them a direct added-value. The lack of an immediate
interest is exactly the eternal problem to which surveys or interviews are subjected. Based upon these ideas, the concept of a dynamic, customizable search interface was developed. Catalogers who work several hours per day have an interest in customizing their search interface to their individual needs. Besides this direct interest towards the users of the tool, it allows to aggregate information on which metadata fields are actually used. The analysis of the log files of the tool can then be used by the collection manager to monitor which fields are used and how their use evolves over time.

Not only the concept, but also the development and implementation of the dynamic search interface within a concrete case study were very much based upon the observations of the first part of the dissertation. In the framework of an existing research collaboration with the development team of CollectiveAccess, it was decided to implement a prototype of the dynamic search interface within this open source collection management software. Section 3.4.1 reported on the dangers of the closed character of proprietary collection registration software and the difficulties that may occur in this context to adjust the software to the local needs of the institution using the software. This, alongside with the extended metadata management functionalities offered by the software, motivated the choice of working with CollectiveAccess.

Apart from the actual tool and software environment, I also needed to find a cultural heritage institution willing to test the prototype by implementing it within its collection registration database. The September 11th Memorial and Museum in New York fulfilled all of the requirements to act as a case study for this chapter. The very recent and evolving nature of its holdings implies that the use of the different metadata fields by the catalogers is also subject to changes, which the log files allow to analyze over time. The museum reflects also in every possible way the contemporary understanding and definition of cultural heritage, as presented in Sections 1.1.3.3 and 2.3 of the first part of the dissertation.

As the programming, implementation and the use of the prototype solely relied on the voluntarism and the belief of its interest to the developers of CollectiveAccess and the September 11th Memorial and Museum, it was not feasible within the context of the dissertation to have the tool operational at an earlier stage. Therefore, the prototype remains very much a work in progress, as I will only have the first statistically significant results based upon an analysis of the log files in a few months.

Independently from the exact future outcomes of the log files, the third operational strategy described in this chapter illustrates in a striking way how the actual use of metadata, subject to the “temps court”, can guide future actions upon the structure of the metadata, analyzed within the “temps intermédiaire”. Section 1.2.3 from the introduction described how the initial one way interaction between the different time levels of Fernand Braudel was extended by Isabelle Boydens. Her work on the quality of
administrative databases demonstrated how changes within the reality of citizens also affect the administration and the legislation of social security. This two way interaction between the levels of the stratified time concept has been confirmed by the prototype developed in this chapter.

Based upon the analysis of the log files, collection managers will have the opportunity to assess the adequacy of the metadata scheme they are currently using to describe and access their holdings. Statistical analyses of the search fields which are added or deleted from customized search forms will point out the relevancy of the individual metadata fields. Collection managers within the cultural heritage sector are very often confronted with back logs of holdings that have not been described due to a lack of resources. The results from the dynamic search interface may give these managers an automated tool to guide their decisions regarding the optimization of the metadata scheme. Based upon the statistical analyses, a collection manager can for example decide to prioritize work on the most used metadata fields and have the initial metadata schema be adapted accordingly.

The first publication on the result of the statistical analysis based upon the log files is foreseen for the summer of 2009. As the dynamic search interface will become a standard feature of the CollectiveAccess software starting from its next major update in the spring of 2009, a significant number of cultural heritage institutions throughout Europe and the United States will start using the dynamic search interface. This will provide supplementary feedback on the direct interest of the interface towards catalogers and after some months also on the use of the logfiles.
Chapter 8

Conclusions

Contents

8.1 Summary of the findings ................................. 182
  8.1.1 Operational strategies ................................. 182
  8.1.2 Critical analysis of the first decade of digitization projects ........... 185
  8.1.3 Holistic approach of metadata quality ......................... 191
  8.1.4 On the importance and the ambiguity of a user-centric metadata quality approach ..................... 196
  8.1.5 Recommendations for metadata practitioners and policy makers .................................................. 199
8.2 Future research ............................................. 203
  8.2.1 Reporting tools for collection registration software 203
  8.2.2 Convergence of libraries, archives and museums 203
  8.2.3 Interaction between digitized cultural heritage and historical research ..................................... 206
  8.2.4 Digital libraries curriculum development ..................... 208

Summary: The conclusions bring together the different research results from the dissertation, in which I have demonstrated that the application of ICT does not always lead to better documentation practices. The outcomes and interests of the three operational strategies are presented, followed by the critical analysis of the first decade of digitization projects. Building on the stratified time concept, the first holistic approach of metadata quality in the cultural heritage sector was developed. The fourth central outcome of the dissertation has demonstrated the importance but also the ambiguity of the relationship between user needs and metadata quality. A condensed set of recommendations and four future research questions complete the conclusions.
8.1 Summary of the findings

Throughout the seven chapters of the dissertation, I have concentrated on how the quality of metadata has evolved in the last decade within the cultural heritage sector, under the influence of digitization projects. As mentioned in the introduction, metadata are throughout the dissertation defined as data about data. Their quality is judged in accordance with the ISO 9000 definition, which refers to the fitness for purpose of an object or a service [79]. The digitization of cultural heritage refers to the remediation of the holdings of libraries, archives and museums from an analogue to a digital medium.

Commonly, the quality of the documentation and the access to our cultural heritage is thought to advance in relation with the explosive growth of computational means. However, the dissertation has proven in detail that this is definitively not the case. Both theoretical developments and numerous examples from case studies demonstrated that the quality of metadata does not augment in a linear way with technological developments. The implementation of ICT for the creation and management of metadata within the cultural heritage sector resembles more the dancing procession of Echternach.1 New technologies clearly offer advantages and need to be implemented within metadata practices, so by no means a ludite vision is propagated within this dissertation. However, as for example the catalog of the city museum of Tienen (illustrated in Figures 1.3 and 1.4 from Section 1.2.2) demonstrated, the quality of the metadata can be affected in a negative sense when technologies are inadequately used. The use of new technologies can be problematic within every application domain, but the specific character of the cultural heritage sector, which was defined in Chapter 2, is responsible for a unique context. The following four points will summarize the principal outcomes of the study of this problem area.

8.1.1 Operational strategies

The second part of the dissertation has concentrated on the development of three operational strategies to act upon the quality of metadata within cultural heritage institutions. Throughout the development of these strategies, I adopted a radical pragmatic approach. Confronted with the limited financial and technical resources of libraries, archives and museums, only low-cost and technically straightforward tools were mobilized. All of the software mentioned within the strategies is open source and available under the GNU General Public License (GPL). Secondly, the different types

1The dancing procession of the city of Echternach in Luxemburg is famous for the way in which the participants of the procession have to take two steps back after every three steps forward. Therefore, five steps are needed to advance one pace.
8.1. SUMMARY OF THE FINDINGS

of software can be implemented and operated by metadata practitioners who do not necessarily need to have a computer science background.

Apart from their practical use toward metadata practitioners, the three strategies complement and reinforce the outcomes of the first part of the dissertation. The continuum between the different levels of the stratified time concept, described in Section 1.2.3, is demonstrated throughout the different operational strategies. The analysis of the metadata themselves and the day-to-day interaction between the metadata and their users deliver outcomes which can be deployed to adapt the “temps intermédiaire” and the “temps long”. The case studies used to illustrate the strategies also reflected the evolutions described in the first part of the dissertation. The use of the September 11th Memorial and Museum as a case study in Chapter 7 for example allowed to back up the assertions made on the evolving character of cultural heritage from Section 1.1.3.3.

8.1.1.1 Data-profiling

Chapter 5 demonstrated how simple and inexpensive data-profiling techniques can be used within the cultural heritage sector to bring a great deal of problems and particularities within large sets of metadata to the surface. The description of large cultural heritage collections is a long process during which different technologies are used by different catalogers. Databases that currently contain ten or hundred thousands of records are the result of several decades of cataloging by different individuals, using evolving rules and tools. While confronted with these large metadata sets, it is difficult to assess their overall quality and coherence. Samples can be manually analyzed, but this method is very resource intensive and only produces a “snapshot” of the quality of metadata at a specific moment in time.

An export of the database from the ethnographic department of the Royal Museum for Central Africa served as a case study to test the outcome of five scripts from a general purpose data-profiler. The high percentage of empty elds is the most striking outcome of the performed analyses. Other detected problems included field overloading and inconsistencies in the syntax of metadata fields, such as the field concerning dates and dimensions. The perception of these issues allows a better understanding of the problems and incoherences within large sets of metadata records and can guide future modifications and developments of the metadata scheme used by collection holders. This illustrates how the short time can have an impact upon the intermediate time.
CHAPTER 8. CONCLUSIONS

8.1.1.2 User comments

The second strategy of the second part of the dissertation analyzed which role user comments can play to augment and correct existing metadata within cultural heritage institutions. Two maximum variation case studies differing on the level of expertise of the user comments, delivered empirical results that point out the possibilities offered by user comments as a pragmatic method to enhance metadata quality. I analyzed and mapped both the content of user queries and comments of the image database of the National Archives of the Netherlands and the Ross Archive of African Images to the Shatford classification. In both cases, queries and comments are highly motivated by interests in specific terms. As a clear relation between the user queries and comments was found in both case studies, I concluded that the user comments are relevant toward the user community.

In a secondary phase, I investigated how many of the user comments effectively propose corrections to and/or enhancements of the existing metadata, by developing for each case study a typology of the user comments. Here I did not rely on a pre-existing typology but developed in a bottom-up approach a new typology for each case study by manually analyzing samples of the comments. The development of the typologies for each case study clearly showed that user comments can play an important role for adding extra metadata or for obtaining feedback on the validity of the existing metadata. Almost half of the comments from the large public and practically all of the comments from the experts fulfilled these roles. The second important outcome from the typologies was the low rate of personal and subjective comments from the general public.

The interaction between users and existing metadata through comments offered another example of how the short time can influence the upper levels of the stratified time. Increasing the interactivity between cultural heritage and its users is considered by policy makers as one of the main advantages of publishing cultural heritage online, as I illustrated in Section 2.3.1. The concrete experiences and analyses of user comments can in their turn influence the development of a long term policy on the issue of user-generated metadata.

8.1.1.3 Dynamic search interface

Chapter 7 provided the third and last strategy for the evaluation of metadata quality. Confronted with the importance but also the difficulties of dening user needs sketched in the first part of the dissertation, an original concept for automatically detecting user needs is developed. With the help of the developers of CollectiveAccess, a dynamic search interface was programmed and implemented within the collection registration database of the September 11th Memorial and Museum. Influenced by the success of
8.1. SUMMARY OF THE FINDINGS

social web applications such as Facebook and Last.fm, the dynamic search interface gives catalogers the possibility to customize their own search interface. By customizing the search fields, the users create log files which are then used to analyze the actual use of the metadata fields.

The prototype was implemented in the museum in December - January 2009 and the first statistically significant results based upon the log files are expected to become available around May - June 2009. However, independently from the exact future outcomes of the log files, the third operational strategy described in Chapter 7 again illustrates in a striking way how the actual use of metadata, subject to the tems court, can guide future actions upon the structure of the metadata, studied in the intermediate time.

8.1.2 Critical analysis of the first decade of digitization projects

After a decade of experiencing digitization projects in the cultural heritage domain, a critical evaluation of the outcomes of these activities is more necessary then ever. There exists a consensus within the sector that the experimentation phase is over and that time has come to scale up digitization initiatives. However, this dissertation has demonstrated that the metadata creation and management within these digitization projects is not a straightforward process.

As we enter the next era of digitization within the cultural heritage domain, in which digitization is now considered an integral part of collection management which has outgrown its experimental character, the problematic areas and possible pitfalls of digitization need to be addressed more explicitly. The dissertation has taken the first step in this direction by paying particular attention to the evolution of metadata quality.

The specific focus on metadata quality allowed me to evaluate the real added value of technologies or standards that have been deployed within digitization projects throughout the last few years. Two issues that are at the crossroads of metadata management and digitization practices have been studied within the dissertation with regard to their impact on metadata quality: the semantic and social web. Both of these recent developments have gained substantial media- and research attention, and are being implemented at varying levels. A critical evaluation of both approaches in the theoretical first part of the dissertation and the outcomes of the three operational strategies from the empirical second part have led us to a third approach: the pragmatic web.
8.1.2.1 On the problem of formalizing low quality metadata

As mentioned in a research paper of the E-Culture project, the semantic web research community strives to facilitate the interoperability between different collections, that would then “become part of one interoperable space” [128]. The essential problem with this approach is that the practical implementation of semantic web standards requires high-quality metadata, which most cultural heritage institutions are unable to offer. Chapter 5 has illustrated the different problems that occur in large sets of metadata records which have been created over several years by various employees. The inconsistent encoding of the date-field that I discovered in Section 5.4.3 as a result of the data-profiling analyses is an obvious example of how real-life metadata are not fit to be formalized to be represented in an interoperable manner.

From a more general point of view, the price that has to be paid for semantic interoperability can be too high within the cultural heritage sector. Within semantic web projects, two opposite operations are performed. All sense of narrative from the initial metadata is broken down by chopping them up in the smallest possible semantic units which can be represented in RDF and OWL. At this stage, new semantics can be automatically inferred by matching the metadata with other metadata record sets, which can be used to enrich them again.

But the question is whether the formalized metadata are still meaningful enough for the human user. The solution toward this problem would be to offer human users the rich, narrative metadata, and have all the highly formalized metadata running in the background on top of the human-readable metadata. At this point we are again confronted with one of the central problems of metadata: they are ever-extendable. Isabelle Boydens has pointed out this danger by referring to Friedrich Nietzsche’s “Vom Nutzen und Nachtheil der Historie fuer das Leben”, in which Nietzsche aims to demonstrate that we need to “serve history only insofar as it serves living” [25, 107]. The different layers of metadata that are added one on top of another result in highly complex and difficult to maintain documentation practices.

The possibility to give comment on a user review of an Amazon product is an illustrative example. The user comment, which can be considered a form of metadata on the book a customer bought, is turned into data itself by attaching metadata to it under the form of a review of the comment. In the context of my specific topic, scheme versioning as described by Joseph Tennis [140] provides a good example of the growing complexity of metadata. He presents a framework to document the changes that occur in a classification scheme or a thesaurus. Another illustrative example is the event-aware metadata model of Carl Lagoze. This model, illustrated in

\[\text{http://records.viu.ca/johnstoi/Nietzsche/history.htm.}\]
8.1. SUMMARY OF THE FINDINGS

Figure 8.1, is interesting on a conceptual level, but its implementation in a real-life application is unlikely, as it adds a complementary complexity to the metadata structure.

Figure 8.1: Event-aware approach toward metadata management

These examples illustrate how different layers of metadata can continuously be added within an information system.

8.1.2.2 Acting in the present instead of describing the past

User-generated metadata, which are a central element of the social web, are currently one of the main points of discussion in the domain of information retrieval on the Internet. I discussed various types of user-generated metadata in different sections of the dissertation and analyzed the opportunities of user-comments in depth with the help of two case studies in Chapter 6.

The most debated type of user-generated metadata are folksonomies, which emerged from the blogging community and social networking websites around 2003-2004. They gained much popularity in the next few years throughout different sectors. Several large scale projects have been launched in the last years to implement social tagging within the websites and collection registration software of cultural heritage institutions.

I gave an overview of the possibilities and issues concerning folksonomies in Section 3.5.2. Here I would like to re-focus on an unacknowledged aspect of folksonomies. Geoffrey Bowker has pointed out that a considerable but overlooked part of our memory practices merely function as practices to manage and frame the present, without having the actual intention of
recalling the past [23]. Folksonomies are a new and interesting example of this theory. Analyzing social tagging merely as a new tool for information retrieval denies the important role folksonomies currently play in how cultural heritage institutions define the relationship with their public. The main function of including for example tagclouds on the homepage of a museum or a library website is not so much to enable a new form of information retrieval, but to communicate about the open and progressive mentality of a heritage institution that wants to emancipate its userbase. The “tag parties” organized in museums, during which the public is invited to add keywords to art objects within the collection registration database, also demonstrate how cultural heritage institutions use this form of metadata creation as a way to define the relationship with their audience. The promotion of folksonomies by funding agencies is another example of how evolutions from the long term, in which I examine shifts within the funding policies of public bodies, influence the actual metadata creation process. Publications such as “E-Cultuur. Bouwstenen voor praktijk en beleid” [160] and “The handbook on cultural web user interaction” [29] are both representing the best practices as defined by public funding agencies. Within both publications, folksonomies and user-generated metadata are explicitly presented as new progressive tools for information retrieval. At no point these publications include any empirical foundation as to how and in which measure folksonomies enhance information retrieval.

Ironically, the adherents of folksonomies claim that social tagging allows an indexation which is culture and world-view independent. Traditional controlled vocabularies such as taxonomies and thesauri are considered to represent specific world views and mentalities and to obstruct a diversity of opinions. However, if we take a closer look at the tagcloud from the social bookmarking site Delicious, which is represented in Figure 8.2, one can hardly state that this search- and browse interface reflects a neutral point of view on the totality of information which is available on the internet.3 The majority of tags are all closely related to new technologies and reflect the interests of a young and western public.

The example of the tagcloud again illustrates the danger of documenting cultural heritage with an interface which reflects in a very direct manner the current mentality. Geoffrey Bowker describes adequately the stakes to which we are confronted:

We need to hold the past open so that we do not hypostasize and freeze the present, and by extension limit our future[23, p. 228].

This quote of Bowker expresses the same thoughts which led Benjamin Barber to work on a broader view of democracy. Section 2.2.3.1 presented how Barber elaborated a broader vision of democracy, by also taking into

---

3See the website of this social bookmarking tool on http://delicious.com
account previous and future generations. By referring to the concept of the "social pact" of Edmund Burke, Barber has pointed out that democracy is a partnership not only between the members of the current generation, but also between those who are dead, and those who will come after us [8].

8.1.2.3 Toward more pragmatism and transparency

By focusing on the topic of metadata quality, I evaluated the effective added-value of new technologies and approaches of documenting cultural heritage. The new approaches that were discussed above are often embedded within a hyped sphere where both the advocates and the critics of these approaches do not base their arguments on empirical evidence. The combination of the historical overview of documentation practices presented in Chapter 3.2 and the operational strategies of the second part
of the dissertation allowed to look beyond the hype of the moment. I can now conclude that in spite of the arrival of promising and powerful technologies, metadata creation and management in the cultural heritage sector remain very resource intensive activities which depend on human intervention.\(^4\)

The advances that have been made in the last decade both on the level of collection registration software and metadata standardization are truly encouraging. All too often policy makers and metadata practitioners alike think that these new and positive developments can be seamlessly implemented within an existing context. The different concrete examples mentioned throughout the dissertation have illustrated that this unfortunately is not the case.

The majority of the databases of cultural heritage institutions are currently populated with metadata records that have been created at different time intervals, by different metadata practitioners. Within the limits of this dissertation, I could only perform data-profiling upon one test collection. But throughout my numerous contacts and effective collaborations with metadata practitioners, I observed that the organizational and financial planning of digitization projects frequently becomes problematic as a result of the disappointing level of metadata quality. Only when projects effectively start migrating and matching existing metadata are the problems diagnosed.

Therefore, a fundamental shift in the thinking on metadata quality is necessary. Instead of assuming a basic good quality level of the metadata, practitioners should be more vigilant and follow up the quality of their metadata records in a more active manner, in order to avoid unfortunate discoveries during the course of a project.

In the absence of standardized methodologies and freely available metadata quality tools, practitioners usually believe that producing information describing the quality of their metadata is too big a step to be taken. The second part of the dissertation has filled this gap with the development of three basic strategies to tackle metadata quality issues in practice.

The use of an open source general purpose data-profiling software has demonstrated how a large set of metadata records can be diagnosed in a quick and straightforward manner by using five simple scripts. The results of these scripts deliver basic information on the syntax and formal characteristics of the metadata, which can lead to the detection of fundamental problems regarding the consistency and coherence of the metadata.

After the analysis of the formal characteristics of metadata, a strategy was presented on how a heritage institution can act upon bad quality meta-

\(^4\)Average figures on the cost of metadata creation are hard to calculate, but the blog “Hanging Together” mentions in a specific post on the digitization of cultural heritage that the price per digitized image can run up to 33 dollar, with the majority of that price being spent on metadata creation (see http://hangingtogether.org/?p=573).
8.1. SUMMARY OF THE FINDINGS

data. As the data-profiling only took the formal quality of metadata into consideration, and the manual correction of metadata by professionals is too expensive and thus not a feasible option, two maximum variation case studies, which differ on the level of expertise of the user comments, have delivered empirical results pointing out the possibilities offered by user comments as a pragmatic method to enhance metadata quality.

The third strategy introduced the idea of proactively monitoring the concrete metadata needs of users. With the help of an open source developer, a prototype was developed which offers the catalogers of a recently founded museum a dynamic search- and display interface in the database. The actions that the catalogers performed upon the prototype delivered statistical information on how and which metadata fields are being effectively used.

The combination of these three strategies finally offers cultural heritage holders both the methodology and the tools to strive toward more transparency and accountability regarding the quality of their metadata.

8.1.3 Holistic approach of metadata quality

The dissertation presented the first truly holistic approach toward metadata quality in the cultural heritage sector. The existing literature on this topic has two major drawbacks: 1) metadata are strictly considered structured digital data that only appeared during the second half of the 1990s, and 2) they are studied as the outcome of an independent activity within a cultural heritage institution, which can be studied in an isolated manner.

This dissertation has developed a more holistic approach by stretching out the analysis of metadata quality over two axes: a horizontal axis which represents time, and a vertical axis which represents the different domains that are directly or indirectly concerned with metadata creation and management.

8.1.3.1 The continuum between traditional documentation practices and metadata

I demonstrated that it is impossible to draw a clear line between the outcomes of traditional documentation practices, such as indexing and cataloging, and what is generally considered metadata. Defining metadata as necessary digital excludes the pre-existing outcomes of manual documentation practices that are represented on paper-based media, such as inventory books or card catalogs. The retro-conversion of legacy documentation challenges this clear separation between traditional paper-based finding aids and metadata. In which stage of the digitization of paper-based catalog cards do they conceptually fall under the category of
metadata: with the scanning of the paper cards, the application of Optical Character Recognition (OCR) techniques to convert the image file to a text document, or when splitting up the poorly or non-structured descriptive content in individual and more structured fields within a database?

This continuum between current and past metadata practices is unacknowledged in the literature. Two excellent books were published in 2008 specifically on my topic of metadata. But neither Foulonneau and Riley [56] nor Marcia Lei Zeng and Jian Qin [163] recognize this continuum. Why is the acknowledgement of the continuum between past and current metadata practices so important? As almost all of our cultural heritage institutions largely predate the advent of databases and digital information storage, they posses paper-based legacy documentation which has been gradually migrated to databases.

The term “legacy metadata” is used in this context to refer to metadata that are managed with the combination of an obsolete data format and medium, consisting of hard- and software. But the content itself remains unaltered from one metadata field that is moved to a field with another name. Automated means such as scripts can be applied to adapt the formal character, such as its syntax, of the content so that it better matches the definition of the new field. A typical example is the splitting up of multiple values that were encoded within the same field in a paper-based interface, in multiple fields within a database to enhance retrieval and metadata management.

However, no true tabula rasa can be made from the historic descriptive practices: they leave their traces in the current collection registration database. As fundamental as the work and research on metadata schemes and standardization is, these schemes can rarely be fully applied as a fresh starting point on the work floor. Legacy metadata impose their historically grown structure.

The rapid evolution of descriptive standards and of the technologies needed to implement new metadata standards also stands completely opposite the very slow pace at which changes can be implemented within collection registration practices in cultural institutions. The acronym YAMS, which stands for “Yet Another Metadata Standard”, illustrates the growing cynical attitude of metadata practitioners on the proliferation of metadata schemes, which development sometimes seems to be an end in itself. Blocking out the background and history of current metadata falsely gives the impression that metadata were developed along with the current metadata schema, which does not explain the problems that might occur with their quality.
8.1. SUMMARY OF THE FINDINGS

8.1.3.2 The stratified time

Apart from enlarging the topic of metadata quality by incorporating pre-1990s documentation practices, I pulled my research question open over a vertical axis by incorporating the different domains that are directly or indirectly concerned with metadata creation and management. Here, the concept of the stratified time from Braudel and Boydens intervenes within the dissertation.

The mobilization of the stratified time canvas has produced a new and more complete understanding of metadata quality by exploring a broader range of influences than normally acknowledged in the literature on metadata quality. To summarize the interest of the framework I can return to the initial case that was presented in the introduction of the dissertation. Section 1.2.2 introduced the central research questions of my work by developing a biography of the catalog of the city museum of Tienen. Figure 8.3 represents how the metadata record of a mortar from this museum has evolved from the first catalog in 1900 to the new metadata format that will be adopted in 2009 by the museum.

This concrete example can be used to resume the different evolutions I positioned within the stratified time. The concept, as defined in Section 1.2, consists of three different time layers, which each interact among one another.
In the long time I positioned the evolution of the mentalities and policies adopted by public administrations toward cultural heritage preservation and access, including the shifts in the perception of what should be considered cultural heritage and how the public relates to that heritage. These policies evolve in a slow and gradual way, but have an important impact on metadata quality in the long run.

Referring to the example of the metadata record from the city museum of Tienen, I can point out two central changes. Until the arrival of the Internet, catalogs and databases were considered internal tools for the employees of an institution. Users submitted their questions to a member of the museum staff that could query the database for them. But ever since the Internet has facilitated the publication of cultural heritage, the public administrators have pushed heritage holders to adapt their descriptive practices to the end-user. An example of this is illustrated in the extended description of the object and the fact that digital images of the mortar will be attached to the metadata record. A second example of a long term evolution which can be derived from this example is the gradual transition from French to Dutch as an indexing language. Until the 1940s and 1950s French was used as the official language throughout different sectors, and was gradually replaced by Dutch. The case study from the Royal Museum for Central Africa in Chapter 5 has briefly illustrated how the use of different languages affects metadata quality.

Within the intermediate time I analyzed the evolution of technologies and standards used for metadata creation. Here we can witness through the example the progress from an unstructured description in 1900 within a paper-based document to a highly marked-up metadata record based upon the Dublin Core metadata standard contained within a web-based collection registration software. As I already mentioned earlier, this example from the intermediate time illustrates in a striking manner the growing importance of the database as a cultural form, which was theorized by Lev Manovich in “The language of new media” [97].

Lastly, the evolution of the metadata themselves were analyzed within the short term. In Figure 8.3 we can see for example that the name of the creator of the mortar has slightly changed from “Petrus Vanden Gheyn” to “Petrus Vandenghein”, which is due to the difference in the interpretation of the name engraved in the mortar by two different indexers. The number 20 from the original description has also disappeared, as this referred to the old storage location that no longer exists. On the other hand, the URL from the website of the museum is a new reality which is now included in the metadata.
8.1. SUMMARY OF THE FINDINGS

8.1.3.3 Interdependence of the different time levels

The importance of the stratified time concept flows mainly out of the possibilities of modeling the interaction between the different time-levels. Mentalities and policies from the long time influence both the development of standards and technologies, and the concrete creation of metadata in our cultural institutions. Section 4.3.2 on the evolution of the understanding of quality cataloging clearly demonstrated how the lowered funding of libraries in the 1980s obliged them to simplify metadata standards. The project-based way of financing digitization projects and the short term character of the employment of metadata practitioners that are hired to work on digitization projects, which was discussed in Section 3.4.2, often results in a lack of involvement and understanding of the collections they are describing.

The impact of ICT and documentation standards upon metadata stood central in Chapter 3. Section 3.4.1 demonstrated for example how the black box character of collection registration software negatively influenced metadata quality. The development of new documentation techniques can even affect the content of the cultural objects that are described. Peter Walsh for example showed that the introduction of photography in the second half of the nineteenth century led to a change of conduct regarding the restoration of art objects. Art works that were damaged or incomplete were once restored and completed with replacement parts. The "post-photographic" museum however started to remove all of the non-original aspects and tried to limit restoration to the absolute minimum. Photographs had thus rendered the "originals themselves so important and valuable that curators could no longer tolerate anything less" [155].

In contrast to the original framework developed by Fernand Braudel, where the interaction consists of a unidirectional impact of the longer sequences upon the shorter time levels, Isabelle Boydens has argued that in the context of information systems, the shorter time levels can also influence the longer sequences. This modification of the stratified time concept has been confirmed by the outcomes of this dissertation. Chapter 7 demonstrated how the day-to-day use of metadata can be used to develop recommendations to modify the structure of the metadata, which are discussed within the intermediate time, and to guide the development of a long-term digitization policy.

8.1.3.4 The historization of metadata practices

The crucial merit of the stratified time concept is that this canvas explicitly proves that metadata practices should be studied and evaluated within a much broader context than they have been until now. In other words, the stratified time canvas allows to "historize" metadata practices. I refer
here to the way Alain Desrosieres delivered an analogue interpretation of “historiciser” in an article on the intertwining history of statistics and the type of government in power. He illustrates how the history of statistics, just like the evolution of metadata practices, is mostly considered a linear, progressive movement, that makes use of new emerging technologies to become increasingly more powerful and complex.

This vision however does not correspond with reality, as the evolution of statistics is as turbulent as the ways of organizing and thinking society: “l’histoire des outils de rationalisation est, malgré ce que cherchent parfois à prétendre les rationalisateurs, aussi tumultueuse et non linéaire que celle des façons de penser la société et celle des politiques visant à agir sur celle-ci” [41]. The histories of economical politics and statistics have stimulated and influenced one another in ways that are rarely recognized:

En un mot, ces deux histoires, celle des politiques économiques, et celle de la statistique, sont rarement présentées et surtout problématisées ensemble. La cause de la quasi-absence de cette histoire dans les travaux sur l’histoire économique est simple. La statistique y est perçue comme un instrument, une méthodologie subordonnée, un outil technique fournissant une validation empirique aux recherches économiques et leurs usages politiques. Dans cette conception linéaire du progrès de la science et de ses applications, la statistique (en tant que production de données et en tant qu’outil mathématique d’analyse de celles-ci) ne peut évoluer que de façon autonome par rapport aux doctrines et aux pratiques économiques. C’est pour cette raison que, dans les ouvrages d’histoire de la pensée ou des faits économiques, cet aspect est peu traité, et n’est, en tous cas, jamais envisagé comme problématique et éventuellement contradictoire, c’est à dire digne d’un développement historique spécifique [41, p. 208].

A clear analogy can be made with how digitization and documentation practices are merely thought of as instruments which evolve in a progressive, linear movement. The dissertation has extensively demonstrated that this is not the case.

8.1.4 On the importance and the ambiguity of a user-centric metadata quality approach

All throughout the dissertation I adopted a user-centric approach toward metadata quality. Based upon the ISO 9000 definition of quality, which is commonly referred to as the “fitness for purpose” criterium (see Section 1.1.3.2), the literature on the quality of metadata, and data more generally, has slowly adopted this vision in the past several years. As
Foulonneau and Riley note, the record-centric approach of metadata quality needs to be replaced by a user-centric perspective: “Think of what a record can do, rather than thinking of it as an artifact in and of itself” [56, p. 56].

But as I demonstrated in Section 2.2, the definition of what exactly user needs are and how they can be measured, has traditionally been a difficult process within the cultural heritage sector. The overview of some concrete user needs assessment reports showed that these evaluation projects can be very tedious and expensive.

However, the digitization of cultural heritage has revolutionized the possibilities for the interaction between a user and the cultural heritage resource he is interested in. This interaction can be used as a new basis from which user needs can be deducted. Two of the strategies applied in the operational part of the dissertation were built upon this interaction facilitated by the Internet and databases. Chapter 6 has evaluated the adequacy of user comments by matching them with user queries. Chapter 7 has taken user interaction as a point of departure to derive user needs. The customization of the search- and display interface by the users of a collection registration database delivered statistics in an automated manner that can be used to evaluate metadata quality on an ongoing basis.

It is interesting at this point to move the discussion regarding user needs to a more general level by debating the impact of the Internet and digitization on the role of the user. A broad range of literature has emerged after the maturation of the Internet, which points to an emancipation of the user on the web. Yochai Benkler describes this as the emergence of a new type of folk culture, which makes “culture more participatory, and renders it more legible to all of its inhabitants” [12, p. 300].

But this liberating and emancipatory effect upon the users is also questioned. Andrew Keene caused a debate with the publication of “Cult of the amateur: how today’s Internet is killing our culture” [82], in which he heavily criticizes the way the Internet is acting as a tribune for mediocrity and amateurism. It is outside the scope of the dissertation to enter this debate, but the ambiguity that accompanies user interaction on the Internet also has its echos within the cultural heritage field.

Although user comments can offer an added-value and user needs must be taken into account when defining and evaluating metadata quality, cultural heritage cannot be placed as a regular player on the market. By attributing an absolute importance to the “fitness for purpose” definition of quality, the demand of the current user regulates in a disturbing manner the offer of our cultural institutions. By referring to the work of Benjamin Barber, I broadened the notion of the user through the accentuation of the importance of the future users, whose needs we cannot necessarily predict.

Ultimately, the dichotomy I observed between the user-comments and the
professionally created metadata in the case study of Chapter 6 should be relativized. Specifically in the context of history and cultural heritage, the apparent tension that exists between metadata produced by experts and laymen does not need to be problematic, and can even be revealed as a necessity.

In her article “Funktionsgedächtnis und Speichergedächtnis. Zwei Modi der Erinnerung”, Aleida Assmann demonstrates that this tension is intrinsically incorporated in the notion of history [7]. She reveals, by referring to the work of Friedrich Nietzsche, Maurice Halbwachs and Pierre Nora, that the apparent dichotomy between “functional” and “factual” memory is artificial in the sense that they depend on another. The functional memory corresponds to the current symbolic or emotive interpretation of a historical fact, where the past is relived by incorporating it into the present. The factual memory refers to the scientific historiography, which solely aims to represent the past, completely independent from the contemporary world.

As Assmann explains, the apparent opposition of both types of memory practices is in fact an interdependency, as they keep another in balance. The purely functional memory distorts the past by altering or even inventing historic events, to let them fit into the glorious nationalist past for example. The strictly factual memory on the other hand loses its relevance by its inability to connect to a public:

In ihrer Verschraenkung steckt ein für beide Seiten heilsames Korretiv. Ein vom Speichergedächtnis abgekoppeltes Funktionsgedächtnis verkommt zum Phantasma, ein vom Funktionsgedächtnis abgekoppeltes Speichergedächtnis verkommt zu einer Masse bedeutungsloser Informationen. So wie das Speichergedächtnis das Funktionsgedächtnis verifizieren, stützen oder korrigieren kann, kann das Funktionsgedächtnis das Speichergedächtnis orientieren und motivieren[7, p. 185].5

The characteristics of digitized cultural heritage collections, such as their adaptability and ease of access, have greatly contributed to placing cultural heritage within the functional memory. The possibility for a heritage institution to interact with its public through user comments, and embedding them alongside the professionally created metadata, is a good illustration of how both factual and functional memory can co-exist and mutually support one another.

5Translation from the author: “Their interdependence has a positive effect for both approaches. Without factual memory, the functional memory drifts away into phantasms, and factual memory without functional memory ends up in a collection of meaningless data. Just like the factual memory can check, back up and correct the functional memory, the functional memory can orient and motivate the factual memory.”
8.1. SUMMARY OF THE FINDINGS

8.1.5 Recommendations for metadata practitioners and policy makers

Within this section, a very succinct overview of the overall lessons deduced from the confrontation with the diverse real-life examples from the dissertation is presented. These points are mentioned throughout the dissertation, but this list presents them in the form of a set of recommendations for metadata practitioners and policy makers:

8.1.5.1 No more digitization projects

After a decade of experimentations and project-based experiences, digitization is here to stay and should as soon as possible be “institutionalized”. Digitization needs to be considered one of the central functions of a cultural heritage, and therefore permanent staff should be specifically dedicated to this task. The short-term character of projects has too many negative consequences in practice. Cultural heritage institutions should develop their own long term digitization strategies, based upon the specific character of their collection and the mission of the institution.

This is currently difficult as individual institutions greatly depend from outside financing for performing digitization activities. In the context of an evaluation of the overall digitization policy of the Royal Museum of Central Africa, I calculated that 38% of the salaries of the museums employees who worked on digitization is financed through external funding [150]. The external funding of digitization mainly occurs in the context of projects, which run over a period of one or two years. This way of working results in an important turnover of the employees of an institution and frequent changes in the type and the object of research.

Another negative outcome is the concentration of project funding in a relative restricted circle of institutions. In a caricatural way, one could speak of a digitization project Mafia. At different levels of project funding, the same phenomenon occurs where a small number of individuals, institutions and enterprises keep on appearing throughout different projects. Veronica Perkins referred in this context to the idea of the “Matthew Principle”, which states that “he that hath, to him shall be given, but he that hath not, from him shall be taken even that which he hath” [117]. This can be specifically applied to the context of project funding at a transnational level. In analogy with European funding in the farming sector, where the funding is mainly attributed to large scale farming companies which outrun the small farmers, only the largest cultural heritage institutions can invest in the drafting of project proposals for European funding. Fortunately, local authorities are increasingly funding small, community-based digitization projects.
8.1.5.2 All metadata have problems

All too often, the good quality of existing metadata is taken for granted when planning digitization projects, which results in unexpected delays and costs when the quality of the metadata does not meet the initial expectations.

This can be particularly problematic in the context of collaborative projects, where metadata from different institutions are aggregated. As each institution has its own interpretation of what quality metadata are, conflicts may rise when metadata from different resources are aggregated.

It is in this context of uncertainty regarding metadata quality that Chapter 5 has focused on data-profiling. The case study from the ethnographic department of the Royal Museum of Central-Africa (RMCA) has demonstrated that this methodology from the business and administrative domain is also of interest to the cultural heritage sector. A set of five simple scripts allowed to discover numerous inconsistencies within the metadata set of the RMCA. Their discovery does not resolve the problems, but the data-profiling results give at least an unbiased view of the formal quality of large metadata sets.

Future research should investigate how the results of data-profiling scripts could be automatically embedded within application profiles. Section 4.4 described the general idea behind application profiles, which give the user specifications regarding the actual implementation and use of a metadata schema. Embedding the outcomes of data-profiling within application profiles could extend their functionality of informing users of which quality they can expect from a metadata repository.

8.1.5.3 Metadata creation was, is and will remain expensive

Despite advances in collection registration software and automated metadata creation, professional metadata creation within the cultural heritage sector remains very expensive. Content-based Image Retrieval (CBIR) has made tremendous advances, and is highly operational in several other application domains. However, its application in the cultural heritage domain currently remains limited. This is in particular due to the so called "semantic gap", which describes the difference that exists between the denotative and the connotative meaning of a cultural heritage resource. Section 2.4.1 has provided more background information on this issue. It is difficult to forecast what future developments will bring, but it is a fact that metadata creation will remain mainly a human, and therefore costly, activity within the next few years.

Each context for the creation of metadata is unique, one can therefore not rely on averages to predict costs. As Foulonneau and Riley suggest, the best approach is to take small but representative samples from the
collections that need to be described and calculate per hour and per item costs \[56\]. The RLG Worksheet for Estimating Digital Reformatting Costs then gives the opportunity to calculate the overall costs for an entire project.\[6\]

Fact is that digitization, and the metadata creation and management it implies, forms an extra activity which requires substantial resources. However, as the overall budgets of cultural heritage institutions are far from being augmented, they are obliged to look for supplementary external funding to finance digitization and metadata creation. This leads us back to one of the first remark from this section regarding the negative effect of the project-based way of working on metadata within the cultural heritage sector.

### 8.1.5.4 Automate usage statistics

User needs and their identification formed one of the central points of focus of the dissertation. Direct interactions with users, either through face-to-face or telephone interviews, or through observation in a usability-testing type scenario, have been the most used methods to gather information on user needs. But the preparation, set-up and analysis of interviews and observations are very resource intensive. These methods are also performed on sample populations and only represent a “snapshot” in time.

The case studies on user comments (see Chapter 6) and the interaction between users and a search interface (see Chapter 7) both demonstrated how the actions undertaken by users can be gathered, analyzed and reused to augment the metadata quality. The analysis of the user comments was based upon a sample population which was manually analyzed, as there are currently no straightforward automated means to match the content of the comments to the Shatford classification. The dynamic search interface presented in Chapter 7 did however maximize automation, by monitoring and deriving benefit in an indirect manner of user actions.

Popular social web applications such as Last.fm and Facebook inspired this approach. The implementation of the dynamic search interface prototype in the collection management database of the September 11th Memorial and Museum demonstrated how an online environment provides possibilities to monitor the behavior and actions of catalogers. Currently cultural heritage institutions do not exploit enough the potential of various types of log files.

The clicking behaviour of users can for example be analyzed with the help of heatmaps that highlight in a visual manner where the visitors of a website click. This kind of information can help web masters reorganizing the content and the navigational structure of their website. But other types of information that are created by the user when interacting with 6This worksheet is available on http://www.oclc.org/programs/ourwork/past/digimtools/RLGWorksheet.pdf.
the website offer even more possibilities. The log files of search queries for example are a very rich source of information. Unfortunately, log files of user actions on the web are still underused. The automated production of monthly logging reports should become a standard feature of collection registration software, which could be used by collection managers to guide and adjust their long-term digitization activities.

8.1.5.5 Be pragmatic instead of striving toward perfection

Within the digital environment, standards, hard- and software evolve so quickly that migration should be considered as an inevitable and imminent process. It is a fallacy to think that scanning originals now at the highest possible quality will create a master file that will prevent to re-scan the original ever again. The costs/benefits analysis to decide which level of quality meets the needs for a feasible budget is particularly hard to make within the context of fast technological developments. Mass digitization projects such as the Google Books project, to which I referred in Section 4.3.3, push the current international trend toward a preference for quantity over quality. OCR techniques offer the opportunity to convert paper-based printed books into digital files, upon which the users can perform full-text searches. This is unfortunately not the case with images. As long as CBIR image retrieval techniques are not operational for the cultural heritage domain, manual metadata creation remains inevitable, and will slow down the process of making digital content accessible.

Metadata practitioners increasingly realize that the complete implementation of the descriptive rules of standardized metadata schemes can be too much of a bottleneck when applied in a local context. The development of application profiles, described in Section 4.4, illustrates this tendency toward more pragmatism. The strict minimum rule currently defended is to propose a mapping of the local metadata schema to an upper-level schema such as Dublin Core. This mapping can then facilitate metadata aggregation on a basic level. Section 3.5.3 described how the integration and re-use of existing metadata within different contexts will become more important in the next few years. By facilitating an external access to the database which holds the metadata and the mapping towards an upper-level scheme, third parties can access these existing metadata and build services around them. This re-use of metadata will not only help justifying the initial investments in metadata creation, but also oblige database managers and metadata practitioners to attribute more importance to the evaluation of the quality of their metadata.
8.2 Future research

8.2.1 Reporting tools for collection registration software

The work on the dynamic search interface presented in Chapter 7 remains very much a work in progress. The first statistical significant log files will become available around the summer of 2009. As the prototype of the dynamic search interface will become a standard feature of the next version of CollectiveAccess, scheduled for release in April 2009, all users of the software will start using the dynamic search interface. This will provide me in the year to come with more feedback on the use of the interface in a variety of cultural heritage institutions.

The collaboration with the development team of CollectiveAccess will therefore continue in the future. On a longer term basis I would like to help within the context of the CollectiveAccess-project with the development of an automated reporting tool. Collection managers should be able to print out automatically generated reports which contain all of the relevant information of both the production of records and their use. These reports should present processed statistics on the status of the collection description activities, user queries and comments, web site traffic, etc. A part of these statistics could also be published on the website of a cultural heritage institution. The Indianapolis Museum of Arts is already doing this in some way, by offering a dashboard (illustrated by Figure 8.4) on the museums’ website, giving an overview of the outcomes of specific activities of the institution.

8.2.2 Convergence of libraries, archives and museums

Throughout the entire dissertation, cultural heritage and cultural heritage institutions are consequently used as terms to describe the object and application domain of my dissertation. These terms are not neutral denominators, in the sense that they reflect the increasingly popular concept of eliminating the boundaries between libraries, archives and museums. Section 1.1.3.3 has introduced the historical development of the term cultural heritage, but more extensive research should be undertaken to question whether this assumption is justified or not.

Despite the popularity of the idea to merge libraries, archives and museums, the reality is more complex than most proponents of this idea proclaim. Some researchers prophetically claim that in the digital context, borders between the three institutions will become superfluous: “Information is the capital of the knowledge bases named museums, archives, bibliographies...”

---

7 For a partial list of the users of the software, see http://www.collectiveaccess.org/index.php?g=about&s=who.
CHAPTER 8. CONCLUSIONS

The boundaries between them will collapse; digital collections will combine and create information with a long time value” [85]. In practice however, we see that these institutions are still managed very differently and have varying methodologies and aims.

I already mentioned the MOAC project in Section 2.2.2. The Encoded Archival Description (EAD) was adopted in the context of this project as a descriptive standard, which is normally exclusively used within the archival sector. The idea was that this approach would offer a better contextual understanding of museums objects. This decision was rather un-successful, in the sense that it led to a situation where “images were buried in EAD finding aids at the item-level, leaving users to drill down from the collection-level, through series-levels, and finally, to item-level descriptions to find images” [61]. EAD was then replaced by METS, which provided more opportunities to discover item-level information and the actual digital images.

It is currently unclear whether this trend is merely hype or if this will effectively alter the boundaries between our libraries, archives and muse-
ums. The most well-known and powerful example in terms of impact was the merger in 2004 of the Public Archives of Canada and the National Library of Canada, which resulted in the Library and Archives Canada. The homepage, illustrated in Figure 8.5, mentions that the institution collects and preserves Canada’s documentary heritage.

Figure 8.5: Homepage of the Library and Archives of Canada, formerly known as the Public Archives of Canada and the National Library of Canada

Apart from this rather explicit example of the physical merger of two institutions, the convergence of library, archives and museum collections remains virtual. Two themes are central to this logic of “convergence”. Firstly, the actual procedures for digitizing content largely use the same technologies, so there is a clear motivation to share experiences and competences between different cultural institutions and to build common support structures to save costs. This can range from building up a consortium to outsource large scale digitization projects to private companies in order to obtain a lower price. The same logic applies to the sharing of hosting facilities for example.

More importantly however, there is a wide consensus on the value of
creating common access points or catalogs, that allow users to search in different collections through one interface. The boundaries between the different heritage institutions are being perceived as artificial from the user point of view. But each type of institution represents a different approach toward cataloging and indexing, as illustrated in the example given above from the MOAC-project where museum metadata were encoded with EAD, which led to disappointing results. The development of the Protocol for Metadata Harvesting from the Open Archives Initiative (OAI-PMH) allows the mapping of specific metadata schemes to Dublin Core, which allows end users to do searches through Dublin Core fields over different collections. An entire set of metadata quality issues occur when aggregating records from different participants, that have been identified by researchers such as Muriel Foulonneau [54, 56] and Besiki Stvilia [134].

Research and practical developments will need to point out in the coming years whether this convergence will effectively break down the barriers between libraries, archives and museums.

8.2.3 Interaction between digitized cultural heritage and historical research

Despite the increasing availability of digitized cultural heritage resources, humanities researchers and more particularly historians have been reluctant to take advantage of these resources and the innovative search possibilities that are built around these digital collections. This situation is increasingly acknowledged and initiatives are taken to promote digital humanities research. The European-funded project Digital Research Infrastructure for the Arts and Humanities (DARIAH) is an excellent example of such an initiative. 8

One of the most illustrative domains where digitization can bring enormous progress is the research which uses historical newspapers as a source. Not only does digitization offer a possibility to preserve the extremely fragile and self-destructive historical newspapers, but the maturation of OCR techniques now allows us to perform full-text searches throughout large collections of scanned newspapers. Until a few years ago, historians were obliged to spend enormous amounts of time to consult these resources within the walls of heritage institutions by manually browsing through microfilmed versions of the newspapers.

The following example illustrates how digitization has altered the way of consulting and analyzing newspapers. Figure 8.6 represents the interface of the cultural heritage database “Het Geheugen van Ieper”, which contains a large selection of digitized newspapers that date from the end

---

8See the project website on http://www.dariah.eu/.
of the 19th and the beginning of the 20th century. Historians that are for example doing research on Emile Vandervelde (1866-1938), who was one of the most important Belgian left-wing politicians at the turn of the 19th and 20th century, can now simply launch a full-text search on his name. When the user clicks on one of the results from a list, he or she automatically gets an overview of a newspaper clipping of the specific article which mentions Vandervelde’s name. In the left-hand upper corner of Figure 8.6, the entirely scanned page from the newspaper can be consulted. Within the same interface, the user can save images of interest to a personal folder and add comments.

![Figure 8.6: Display of a search result from a database of digitized historical newspapers](image1)

It is clear that the digitization of cultural heritage can offer new and exciting opportunities for historical research, although it cannot be forgotten that nowadays only a tiny part of our cultural heritage collections are digitally available. In the New York Times article “History, Digitized (and Abridged)” one correctly points to the increasing danger of only considering the digitally available collections as a part of our cultural canon [69]. But methodologies from the history domain can also deliver a much needed framework to deal with current problems such as the authenticity of digital documents. Isabelle Boydens has demonstrated within her monography “Informatique, normes et temps” the force of hermeneutics and

---

9See http://www.geheugenvanieper.be/.
the concept of the stratified time to evaluate the quality of administrative databases [25]. She demonstrated how the monitoring of anomalies within databases, that appear in the short time, can help to punctually update the structure of the database, which is analyzed within the intermediate time. The three operational strategies from the second part of this dissertation have clearly shown that the stratified time can also be applied within the specific context of cultural heritage metadata. More work needs to be done to fully develop the power of the framework within digital humanities.

8.2.4 Digital libraries curriculum development

This dissertation has focused on different elements which have created a shift in the thinking and practice of metadata creation and management. Within the framework of the dissertation, all of these elements are studied to analyze metadata quality. However, the evolution of the professional status of metadata practitioners merits a profound analysis in and of itself. Section 3.4.2 delivers an introductory study on the issue, but more empirical and theoretical research needs to be undertaken to grasp the effective changes of the metadata "praxis".

The de-professionalisation I briefly mentioned in Section 3.3 is one of the crucial characteristics and outcomes of this evolution, specifically in the creation of metadata. It is increasingly performed by young and unexperienced employees. One of the central problems lies in the low status of this professional activity. As Bowker notes, "the general problem remains that across the sciences, the activity of naming is mundane and has a low status, even though it is an activity central to the development of a good database" [23].

Two possible long-term strategies can be considered to counter this evolution. Firstly the status of documentation practices could be augmented by incorporating metadata management activities on a scientific curriculum vitae. When performing an assessment of the digitization projects from the different departments of the Royal Museum of Central Africa, several senior scientists mentioned during interviews that they did not wish to invest too much time on the quality control of metadata, as they received no official recognition for this type of work professionally [150]. Therefore, metadata creation and management should be considered as a new kind of scientific publication, that could be included and valorized within an scientific curriculum vitae. But this change in policy also asks for standardized methodologies to evaluate the quality of metadata.

But on a more fundamental level, we also need to evaluate and re-think how the current education of future metadata practitioners and managers can be adapted to the quickly evolving technologies and practices within public and private institutions. Foulonneau and Riley provide a good
overview of the very heterogeneous competencies a metadata specialist should possess [56]. However, the curricula within the Information and Library schools currently do not reflect the study of these diverse competencies.

Confronted with this discrepancy between the content of the education and the required competencies on the work floor, Information and Library Science researchers are currently reflecting on how they can adapt their teaching programs and curriculums. From 2004 onwards, a steadily growing number of publications appeared that reflect specifically on this issue, by concentrating on the content of “Digital Libraries” education.10

In the autumn of 2008, a workshop on this topic was organized to bring together a number of professionals and researchers from Information and Library Science schools in the United States and Europe.11 The project partners plan to gradually work toward a common curriculum, which would allow us to bring the specific competencies from American and European universities together.

A first step in this direction will be the development of an e-learning environment where teachers and students from the participating institutions can collaboratively use open source digital library applications for hands-on experience with the development and use of metadata schemas.

---

10 See for example [119].
11 For an overview of the goals and the attendees of this workshop, please see http://www.ibiblio.org/pomerantz/wiki/index.php/NSF_Workshop_Coordinating_EU-US_Digital_Library_Education.
Appendices
Appendix A

Primary sources

Apart from the case studies presented in the dissertation, other primary sources in the form of interviews helped developing the research questions of this dissertation.\(^1\) The direct contact through interviews with metadata practitioners, policy makers and consultants provided an extremely valuable source of information.

A.1 Interviews

Throughout the period of the dissertation, researchers and professionals active within the cultural heritage sector provided input on past and current digitization projects and feedback on the research questions of the dissertation. My contacts and discussions with collection holders and the managers of digitization projects mostly occurred in an informal manner during meetings, conferences or visits to institutions.

I mainly interviewed the responsible employees for metadata management, to discuss the problems occurring within their database, and asked for examples to be used within this dissertation to illustrate specific problems or issues. This is for example the case with the University Library of Ghent. In other cases, such as the city museum of Tienen I actively participated in the process of defining the metadata conversion process and planning the implementation of a new collection management system. The evaluation of the digitization policy throughout the different scientific departments of the Royal Museum for Central Africa provided me an extensive overview of the organizational and financial implications of digitization projects. The work as a project evaluator for different funding agencies such as the European Commission and the Region of Brussels

\(^1\)In contrast with secondary sources, that is the literature represented in the bibliography, primary sources deliver a direct, unmediated contact with the research topic.
provided me with sufficient knowledge of public administration policies toward digitization projects. Throughout these different contexts, I have interviewed the following people:

– Metadata practitioners:

- Erik Buelinckx, Institut Royal du Patrimoine de l’Art (Brussels, Belgium). Project manager of the image database.
- Christophe Delhaise, Accès Informatisé aux Collections des Institutions Muséales (Namur, Belgium). Consultant.
- Pierre-Yves Desaive, Musées royaux des Beaux-Arts (Brussels, Belgium). Head of the collection registration database.
- Dirk de Wit. Instituut voor beeldende, audiovisuele en mediakunst (Ghent, Belgium). Director.
- Adena Langer, National September 11th Memorial and Museum (New York, United States). Memorial exhibition manager and curatorial assistant.
- Jef Malliet, Provincie Limburg (Hasselt, Belgium). Project manager.
- Dries Moreels, Vlaams Theater Instituut (Brussels, Belgium). Head of collections.
- Hendrik Olivier, Archives and Museum of the Socialist Labour party (Ghent, Belgium). Head of the audiovisual department.
- Richard Rinehart, UC Berkeley Art Museum/Pacific Film Archive (California, United States). Digital media director and adjunct curator.
- Saskia Scheltjens, University Library Ghent (Ghent, Brussels). Head of the humanities library.
- Vincent Seghers, Digipolis (Ghent, Belgium). ICT project manager.
- Liesbeth Thiers, Erfgoedcel Ieper (Ypres, Belgium). Project officer.
- Hein Vanhee, Royal Museum of Central Africa (Tervuren, Belgium). Head of collections.
- Astrid Van Kersschaever, University Library Ghent (Ghent, Brussels). Head of cataloging.
- Jan Vermassen, Art in Flanders (Ghent, Belgium). Director.

– Policy makers:

- Debbie Esmans, Flemish Ministry of Culture (Brussels, Belgium). Project manager.
- Cécile Hubaut, Service du patrimoine culturel du ministère de la Communauté française (Brussels, Belgium). Project manager.
A.1. INTERVIEWS

* Patricia Manson, Directorate General Information Society and Media of the European Commission (Luxembourg). Head of unit.

– **Consultants:**

* Seth Kaufman, Whirl-i-Gig (New York, United States). Director.
* Peter Schouten, Ingressus (Rotterdam, The Netherlands). Consultant.
* Onno Zaman, Pictura Database Publishing (Heiloo, The Netherlands). Director.
Appendix B

Secondary sources

Other secondary sources that do not consist of published scientific literature, such as websites, blogs and mailings lists are represented in this overview. These electronic online resources were lastly consulted in January 2009.

B.1 Websites

The following selection of websites of organizations have been consulted throughout the creation of the dissertation, to obtain the most updated information on metadata standards and technologies.

- **Dublin Core**: engaged in the development of interoperable online metadata standards (http://dublincore.org).
- **World Wide Web Consortium (W3C)**: develops interoperable technologies for the web (http://www.w3.org).
- **Museum Computer Network (MCN)**: supports museums by providing opportunities to explore and disseminate new technologies and best practices (http://www.mcn.edu).
- **Online Computer Library Center Programs and Research (OCLC RLG)**: supports exploration, innovation and community building on behalf of libraries, archives and museums (http://www.oclc.org).
- **Institute of Museum and Library Services (IMLS)**: provides federal support for the American libraries and museums (http://www.imls.gov).
- **Data Archiving and Networked Services (DANS)**: disseminates research results for the long-term storage of digital data from the natural and human sciences (http://www.dans.knaw.nl).
APPENDIX B. SECUNDARY SOURCES

- **Faro. Vlaams steunpunt voor cultureel erfgoed**: provides support and networking activities for the Flemish cultural heritage sector (http://www.faronet.be).

### B.2 Blogs

The RSS feeds of blogs on the specific topic of metadata and/or digitization projects in the cultural heritage sector provided personal views of experts and practitioners.

- **Digitization Blog**: Mark Jordan focuses on digitization and related activities in libraries, archives, and museums, and provides news relevant to people who manage and implement digitization projects (http://digitizationblog.interoperating.info).
- **Hanging Together**: maintained by the staff at RLG Programs which investigates the partnerships between libraries, archives, and museums, and discusses the intersections between these three different types of institutions (http://hangingtogether.org).
- **Les Petites Cases**: Gautier Poupeau discusses the French perspective on developments from the semantic web community for the humanities (http://www.lespetitescases.net).
- **Lorcan Dempsey's weblog**: Lorcan Dempsey publishes research on libraries, services and networks (http://orweblog.oclc.org).
- **Metadata matters**: Diane Hillmann presents research results on this blog (http://managemetadata.org/blog).
- **Musematic**: Gives the latest trends in the world of museum informatics and technology. Experts from the Museum Computer Network and AAM's Media Technology Committee such as Holly Withney and Amalyah Keshet post on this blog (http://musematic.net).
- **Museotech**: Blog on museum technology and Web 2.0. by Jim Angus and Bruce Falk (http://blog.museotech.net).
- **The Doofer Call**: Research diary of Jeremy Ottevanger about museums and the semantic web (http://doofercall.blogspot.com).
- **E-cultuur**: Blog on e-culture in Flanders, on which invited experts post their recommendations and research findings (http://www.e-cultuur.be).

### B.3 Mailing lists

These virtual public places offer the opportunity to follow and participate in discussions regarding the most recent developments and allow a member of a specific community to publish questions or opinions.
B.3. MAILING LISTS

- Museum Computer Network (MCN) mailing list: International group, although mainly US-based participants, of cultural heritage practitioners who work with ICT.
- Dublin Core General mailing list: International group of metadata practitioners and researchers.
- Semantic web W3C mailing list: International group of semantic web researchers.
- Nedlib: Dutch information professionals
- Canmuse-l: Bilingual mailing list for the Canadian heritage community.
Bibliography


[148] M. van der Graaf. De digitale dienstverlening van de uba en de verdere ontwikkeling daarvan: het gebruikersperspectief eindrapport over het gebruikersonderzoek over de digitale bibliotheek onder uva studenten en medewerkers d.m.v. focusgroepen en een online enquête. Consultancy report written for the university library of Amsterdam, 2008.


