Lars Fredrik Høimyr Edvardsen

Using the structural content of documents to automatically generate quality metadata

Thesis for the degree of Philosophiae Doctor

Trondheim, February 2013

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics
and Electrical Engineering
Department of Computer and Information Science

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Computer and Information Science

ISBN 978-82-471-4212-7 (printed ver.)
ISBN 978-82-471-4213-4 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2013:58

Printed by NTNU-trykk

Abstract

During the last decades, document sharing has become vastly more available for the general public, with large document collections being made generally available on the internet and inside of organizations on intranets. In addition, each of us has an ever-increasing archive of private digital documents. At the same time efforts to enable more efficient document retrieval have only succeeded marginally. This makes finding the right document like looking for a needle in the haystack. Just now it is a bigger haystack. This lack of overview of existing document resources results in large amounts of scarce human resources that are still being used to create similar resources.

A key reason to why we are faced with this challenge is that few documents receive a sufficient metadata description in order to enable efficient retrieval. Too often the document metadata is insufficient or even incorrect. Few document creators are aware of describing their documents with metadata. Trained librarians and archivists can assist authors to create and publish metadata, but this is a costly and time-consuming process. Advanced metadata formats, such as the IEEE LOM, enable detailed and precise metadata descriptions. This format is challenging to use and the potential in the format is often not leveraged. Document formats that require such metadata, e.g. SCORM Learning Objects (LOs), are not being used to their potential due to the challenges of creating metadata.

This thesis shows how Automatic Metadata Generation (AMG) can stand as a foundation for creation, publishing and discovery of document resources with rich and correct metadata descriptions. This thesis shows how high quality metadata can be created automatically using the documents themselves and contextual data sources. Finally, this thesis shows how metadata descriptions can be used alongside the original document to create SCORM LOs to enable sharing of educational resources with educational metadata descriptions.

The main contributions by this thesis are:

C1: Establishing an overview of research literature, projects and products using AMG and the quality of their generated metadata.

C2: Establishing that AMG efforts can be combined to expand the range of elements and entities that can be generated, but also to increase the quality of generated entities.

C3: Establishing that AMG efforts can generate high quality metadata from non-homogeneous document collections, vastly expanding the practical usefulness of AMG.

C4: Establishing that AMG efforts can contribute extensively in promoting sharing of knowledge with the creation of sharable SCORM LOs containing the educational resources themselves and extensive metadata descriptions to enable efficient location and use.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfilment of the requirements for the degree of philosophiae doctor.

This doctoral work has been performed at the Department of Computer and Information Science, NTNU, Trondheim, with Ingeborg Torvik Sølvberg and with co-supervisors Trond Aalberg and Hallvard Trætteberg.

Acknowledgements

This research would not be possible to execute without the support of many people. The thesis author would like to use this opportunity to thank everyone involved, directly or indirectly by allowing for creation of a dataset to perform the studies upon.

Supervisor: Ingeborg Torvik Sølvberg

Co-supervisors: Trond Aalberg and Hallvard Trætteberg.

Miscellaneous advisors: Leif Martin Hokstad, Trond Håvard Hanssen, Bård Kjos, Marte Bratseth Johansen, Kristin H Andersen and Kjell Bratbergsengen.

Persons who gave this thesis access to subjects at the case LMS: Anders Sanne, Anne Sigrid Nordby, Arild Holm Clausen, Arne Kristian Myhre, Arne Aalberg, Baggerud Bjørn, Berit Bungum, Bojana Gajic, Curtis H. Whitson, Dankert Vedeler, Guttorm Sindre, Heie Aage Christen, Helge Klungland, Jan Tro, Jens Oluf Andersen, Jon Andreas Støvneng, Karl Vincent Høiseth, Kjell Atle Halvorsen, Lars Grande, Lasse Natvig, Magnus Rasmussen, Marit Støre Valen, Monica Divitini, Nora Hermansen, Novakovic Vojislav, Ola Hunderi, Pauline Haddow, Per Christian Hestbek, Per Gunnar Kjeldsberg, Preben C. Mørk, Sægrov Sveinung, Tommy Gravdahl, Tor G. Syvertsen, Tor Ramstad, Torbjørn Skramstad, Trine Elaine Eiken, Tore Undeland, Knut Einar Larsen, Anne Fiksdahl, Leif Rune Hellevik, Arne Valberg, Catharina Davies, Tore H. Løvaas, Magne Runde, Lisbeth Aune, Sverre Steen, Signe Kjelstrup, Kolbein Bell, Håkan Hytteborn, Bente Sellereite, Bjørn E. Christensen, Svein Valla, Claus Bech, Terje Malvik, Turid Rustad, Gustav Erik Gullikstad Karlsaune, Vera Sandlund, Geir Hansen, Ivar Berthling, Per Bruheim, Espen Robstad Jakobsen, Øystein Røkke, Håkon Kvåle Gissinger, Rudolf Schmid, Amund Bruland and Elin Tonset.

People who gave other feedback regarding the case LMS: Reidar Conradi, Stig Berge, Jon Andreas Støvneng, Arne Mikkelsen, Tor Ramstad, Michael Kachelriess, Ola Hunderi, Jorunn Reitan, Leif Edward Ottesen Kennair, Brynjulf Owren, Jon Steinar Gudmundsson, Arne Mikkelsen and Brynjulf Owren.

Contents

Abstract ..... **i**
Preface ..... **iii**
Acknowledgements ..... **v**
Contents ..... **vi**
List of Figures ..... **viii**
List of Tables ..... **x**
Abbreviations and key terms ..... **xi**
1. Introduction **1**

1.1 Motivation *1*
1.2 Problem Outline *4*
1.3 Research Context *5*
1.4 Research Questions *6*
1.5 Research Objectives *6*
1.6 Research Design *7*
1.7 Papers *9*
1.8 Contributions *12*
1.9 Thesis Structure *14*

2. State of the Art **16**

2.1 Defining "Metadata" *16*
2.2 Defining "Quality" and "Metadata Quality" *17*
2.3 Automatic Metadata Generation *19*

2.3.1 Data sources for Automatic Metadata Generation 20
2.3.2 Approaches for Automatic Metadata Generation 32
2.3.3 Development of AMG rules 44
2.3.4 Conflict handling and trust 46
2.3.5 Projects and systems using AMG 47
2.3.6 Summary 57

2.4 Learning Objects *59*

2.4.1 Introduction 59
2.4.2 Defining "Learning Object" 61
2.4.3 Learning Object Metadata schemas 63
2.4.4 Learning Objects and creation of Learning Object Metadata 75
2.4.5 It's learning - The NTNU LMS Intranet 76

3. Context and Research Design **78**

3.1 Reaching the Research Goal *78*
3.2 Research Process *80*

4. Research results **83**

4.1 The process of creating educational documents *84*

4.1.1 Different document creation user environments 85
4.1.2 Creating documents in the system controlled environment..... 87
4.1.3 Creating documents in a user controlled environment 93
4.1.4 Summary..... 100

4.2 Quantitative element analysis *101*

4.2.1 Uploaded stand-alone documents as part of system documents..... 101
4.2.2 The final dataset..... 111
4.2.3 Quantitative Summary..... 117

4.3 Qualitative element analysis *117*

4.3.1 The "Characters," "Words," "Pages" and "Slides" elements..... 118
4.3.2 The "Creator" element..... 130
4.3.3 The "Title" element 135
4.3.4 The "General. Language" element 151
4.3.5 Qualitative Summary..... 155

5. Conclusion, contributions, objectives and future work 156

5.1 Conclusion *156*
5.2 Major Contributions *157*
5.3 Reaching the Objectives *158*
5.4 Recommendations *160*
5.5 Future work *161*
5.6 Concluding Remarks *162*

6. References..... 163
Appendix A: Papers..... 175

P1: Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment *176*
P2: Automatically generating high quality metadata by analyzing the document code of common file types *188*
P3: Using the structural content of documents to automatically generate quality metadata *208*
P4: Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing? *226*
P5: Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects *239*
P6: Creating Metadata is a Costly Manual Process – And it can be Automated *253*

Appendix B: Secondary papers 262

SP1: Use of It's learning at NTNU – A Quantitative and Qualitative study *263*
SP2: Using Document Code to automatically generate high quality metadata: An Auditing case study *266*

List of Figures

Figure 1: Timeline of main research focus and papers over time 9
Figure 2: Data sources and related AMG analysis approaches 21
Figure 3: Increasingly specific levels of context data 23
Figure 4: Use of the value space for "Lifecycle. Contribute. Role" based on Friesen [52]..... 23
Figure 5: Binary document code for a Word document interpreted as text 29
Figure 6: Document code for a PDF document..... 30
Figure 7: Document code for an HTML document..... 31
Figure 8: AMG content analysis algorithms and their data sources..... 33
Figure 9: Embedded metadata from a Word document stored as HTML 34
Figure 10: Embedded metadata from a Word document stored as XML..... 34
Figure 11: Visual characteristics of document content 34
Figure 12: Natural language of document content 34
Figure 13: Document code of document content..... 34
Figure 14: Visual characteristics of a paper 37
Figure 15: Official NTNU letter template in Word format 42
Figure 16: Open XML document code of Figure 15 once filled in..... 42
Figure 17: Example of rules for rule-based algorithms..... 45
Figure 18: Degree of trust in AMG by metadata experts, from Greenberg et al. [60]. .. 47
Figure 19: The Dublin Core schema 64
Figure 20: IEEE LOM (Draft 8)..... 68
Figure 21: The ADN metadata schema 70
Figure 22: The ADN metadata schema (continued from Figure 21)..... 71
Figure 23: A SCORM LO imported into the NTNU LMS 73
Figure 24: Documents from a system controlled environment as data source for AMG efforts..... 85
Figure 25: Converted stand-alone documents as a data source for AMG..... 86
Figure 26: Creating a new document in a system controlled environment (stage 1) 88
Figure 27: Percentage of use of document types in the case LMS..... 89
Figure 28: Creating a new document in a system controlled environment (stage 2) 92
Figure 29: It's learning template for Exercise document 92
Figure 30: It's learning template for Link document 92
Figure 31: Uploading stand-alone documents to an existing system document..... 93
Figure 32: Blank Word template 94
Figure 33: Blank PowerPoint template..... 94
Figure 34: NTNU lecture slide PowerPoint template..... 94
Figure 35: NTNU thesis PowerPoint template..... 94
Figure 36: Creating a new stand-alone document 95
Figure 37: Saving a new document 97
Figure 38: Editing an existing stand-alone document 97
Figure 39: Re-saving an existing stand-alone document..... 98
Figure 40: Converting a previously saved document 99
Figure 41: The pre-study stand-alone document format types (number of files for each document format)..... 102

Figure 42: Similarity between "Title" element candidates (PDF, Word, PowerPoint and Excel document formats) 104
Figure 43: Stand-alone document format types (number of documents for each document format) 112
Figure 44: Stand-alone document types based on the document formats' primary usage area and Dublin Core “Types” 112
Figure 45: Number of metadata elements collected per stand-alone document 114
Figure 46: Example of a multi-page document with 6 logical pages on one technical page 118
Figure 47: Number of Characters (Word documents) 121
Figure 48: Number of words (Word documents) 122
Figure 49: Number of words (PowerPoint documents) 123
Figure 50: Example of PowerPoint document 123
Figure 51: First slide of a document with extreme results 124
Figure 52: Difference between the logical and technical number of pages 126
Figure 53: Comparing the "Pages" element with the visually correct number of pages 127
Figure 54: Verified correctness of embedded creator metadata 132
Figure 55: Verifiable publisher is document creator 134
Figure 56: Distinction between the first visual line and the first line recorded 139
Figure 57: Example of a Word document with a visible title 140
Figure 58: Example of a Word document with alternative visual presentation 140
Figure 59: Heading of the example document stored as a separate XML-file 141
Figure 60: The Open XML document code of the example document. 141
Figure 61: PowerPoint slide with text boxes, images and groups of content 143
Figure 62: Placement of the slide content in the “slide.xml” file. 143
Figure 63: Word document with a spreadsheet and visual “Creator” element as course name 143
Figure 64: PowerPoint document with multiple types of content in a single text section 143
Figure 65: Template with title and sub-title sections 145
Figure 66: XML document code for Figure 67 145
Figure 67: The template in Figure 65 in use 145
Figure 68: Logical structure of algorithm A 148
Figure 69: Logical structure of algorithm B 148
Figure 70: Logical structure of algorithm C 149

List of Tables

Table 1: Relations among Research Questions, Contributions and Papers 14
Table 2: Using local context to create global metadata 24
Table 3: Common document formats 28
Table 4: Taxonomy of Learning Object Types, based on [133] 62
Table 5: Timeline of educational initiatives and standards 65
Table 6: LMS document types 91
Table 7: Common stand-alone document metadata elements 103
Table 8: Documents described in the LMS 106
Table 9: Elements available in the different document formats 110
Table 10: Recorded elements 113
Table 11: Number of elements per stand-alone document format 114
Table 12: Identifiers within stand-alone documents (both datasets) 117
Table 13: “Page” and “Slides” elements 119
Table 14: Issues affecting counting algorithms 121
Table 15: Entities collected from Figure 50 123
Table 16: Entities collected from Figure 51 124
Table 17: Creator metadata from PDF, Word and PowerPoint documents 131
Table 18: Verifiable correct and false embedded creator elements 132
Table 19: Algorithms for generating "Creator" entities based on visual characteristics 133
Table 20: Formatting information available from PDF, Word and PowerPoint documents 133
Table 21: Verifiable publisher as document creator 134
Table 22: Stricter verification of publisher, including multiple authors 135
Table 23: Rule set for the largest font AMG baseline approach 137
Table 24: Results of baseline AMG Title algorithms: Word documents 138
Table 25: Results of baseline AMG Title algorithms: PowerPoint documents 138
Table 26: Formatting of the first three text sections of the Word document example . 142
Table 27: Results of using style tag formatting 146
Table 28: Comparing rules based on visual characteristics and the document code 147
Table 29: Results of advanced AMG Title algorithms: Word documents 150
Table 30: Results of advanced AMG Title algorithms: PowerPoint documents 150

Abbreviations and key terms

Term	Abbreviation or description
ADEPT	Alexandria Digital Earth ProtoType
ADL	Advanced Distributed Learning Initiative
AND	ADEPT, DLESE, NASA
AI	Artificial Intelligence
AMeGA	Automatic Metadata Generation Applications project
AMG	Automatic Metadata Generation
Composite document structure analysis	Heritage of metadata from related documents rather than other environmental sources.
CRF	Conditional Random Fields
DC	Dublin Core
DC-ed	DC educational
DDC	Dewey Decimal Classification system
DLESE	Digital Library for Earth System Education
Document content analysis	Analysis of the document itself in order to generate the metadata
Document context analysis	Analysis of the environment surrounding the document in order to generate metadata.
Document usage analysis	Retrieval of information of actual document usage in order to generate metadata.
DPI	Dots Per Inch
DTD	Document Type Definition
DVHMM	Dual and Variable Hidden Markov Model
Element set	The collection of the metadata elements that describe a document is known as a metadata element set
Embedded metadata	Metadata which is present within a document.
EXIF	Exchangeable image file format
Extraction	The process by which AMG algorithms create metadata that has previously not existed
FRBR	Functional Requirements for Bibliographic Records
GEM	Gateway to Education Materials
GESTALT	Getting Educational Systems Talking Across Leading-Edge Technologies
GPS	Global Positioning System
GUI	Graphical User Interface
Harvesting	Collecting embedded metadata
HMM	Hidden Markov Models
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IFLA	International Federation of Library Associations and Institutions

JPEG	Joint Photographic Experts Group
Learning Object	Any entity, digital or non-digital, that may be used for learning, education, or training.
LMS	Learning Management System
LO	Learning Object
LOM	Learning Object Metadata
LOMGen	Learning Object Metadata Generator (project)
LOR	Learning Object based Repositories
LRE AP	Learning Document Exchange Metadata application profile
LTSC	(IEEE) Learning Technology Standards Committee
Machine learning algorithms	Rules created by computers for the task of creating rules and determining the weights applied to each rule.
MAGIC	Metadata Automated Generation for Instructional Content
Metadata	Metadata, or structured data about data, improves discovery of and access to such information. The effective use of metadata among applications, however, requires common conventions about semantics, syntax, and structure. Individual resource description communities define the semantics, or meaning, of metadata that address their particular needs. Syntax, the systematic arrangement of data elements for machine-processing, facilitates the exchange and use of metadata among multiple applications. Structure can be thought of as a formal constraint on the syntax for the consistent representation of semantics.
Metadata Quality	They categorized "metadata quality" into seven categories: (1) Completeness, (2) Accuracy, (3) Provenance, (4) Conformance to expectations, (5) Logical consistency and coherence, (6) Timeliness and (7) Accessibility.
Metadata schema	The metadata schema is a systematic and orderly combination of elements used to specify valid element types, the entities that they can contain, and how these element types can be used.
MGS	Metadata Generation System
MIME	Multipurpose Internet Mail Extensions
MS	Microsoft
NASA	National Aeronautics and Space Administration
Natural language	Language of the visual characteristics of the document presents.
NSDL	National Science Digital Library
NTNU	Norwegian University of Science and Technology
OCR	Optical Character Recognition
PDF	Portable Document Format
Quality	Term based on the three category types: (1) Syntax: Analysis of the document formatting to see that it complies with the document's format standard. (2) Semantics: Analysis of the

	entities presented to see if they are valid and in accordance with the document format's relevant metadata schema. (3) Pragmatics: Analysis to determine if the user-interpreted properties are reflected in the metadata.
Rule-based algorithm	Pre-defined manually created rules.
SCORM	Sharable Content Object Reference Model
Source Code	Any sequence of statements and/or declarations that are written in some human-readable computer programming language.
SVM	Support Vector Machine
URL	Uniform resource locator
UUID	Universally Unique Identifier
XMP	Extensible Metadata Platform

1. Introduction

1.1 Motivation

Large amounts of scarce human resources are still being used to create similar resources, such as documents [65]; partly because people are not aware of others' work through lack of sharing opportunities, and the inability to retrieve available documents. During the last decades, document sharing has become vastly more available for the general public, with large document collections being made generally available on the internet and inside of organizations on intranets. In addition, each of us has an ever-increasing archive of private digital documents. At the same time efforts to enable more efficient document retrieval have only evolved marginally. This makes finding *the* right document like looking for a needle in the haystack. Just now it is a bigger haystack. The challenge consists of three factors:

- Describing documents accurately so that the querying user can receive the information required to distinguish documents
- Describing documents accurately so that the search engine can perform an accurate query based on the user presented input, and
- Promotion of document characteristics so that the querying user understands that the promoted document is the desired document

Search engines have grasped the challenge of locating large amounts of documents and promotion of a set of standard characteristics to the querying user through the query results. General purpose search engines commonly promote descriptive characteristics such as a document title, some document body text, a last edited date and document location. Scientific and other purpose specific search engines often also promote descriptions including the document author, keywords and subject. The search engines hence rely upon an accurate document description in order to perform as desired.

This brings us down to how document descriptions are to be created. Some search engines rely on computer programs to identify the document characteristics. Others base their efforts on human created document descriptions. Both methods face considerable challenges.

- With computer program created document descriptions, the accuracy of the descriptions are commonly low, documented in Chapter 2.3.5. In addition, only a few descriptive characteristics are registered, making accurate querying impossible.
- With manually created document descriptions, the number of described documents is limited by humans' time and ability. In addition, there is the issue of human errors and inconsistency which can reduce the accuracy of the document descriptions.

Manual generation of accurate document descriptions, or "quality metadata", requires time and skilled human resources. Research has also shown that the general public also

has little willingness to manually create metadata [23, 27]. As a result, few documents receive high quality metadata descriptions:

> *Manual creation of metadata is tedious, error-prone and doesn't scale. As the amount of learning content continues to grow, it becomes less and less feasible to describe all available content manually. Moreover, the metadata humans create are not perfect. Therefore, we need a change in approach, trying to automate this process as much as possible.*
>
> Meire et al. [99]

The number of documents that need a good description is ever-increasing. And we cannot rely on manual efforts to create all the needed document descriptions.

From a scalability perspective, usage of computer programs to create document descriptions is the only viable solution. Though, the quality of the generated descriptions must be brought to a completely new level.

- The generated metadata must be more accurate, so that both the search engines and the querying users can rely upon the available information. And,
- The metadata descriptions must be richer and more detailed, so that the querying users can state more accurately what he or she is looking for and for search engines to present more of the vital information needed for the user to make the optimal document choice.

Many computer programs have been created that create metadata. In addition, there are a number of logical possibilities that has yet to be explored. There is a need to systematically review how metadata can be created using computer programs, in order to achieve creation of high quality document metadata.

There is a need to establish what to do and what not to do with our computer programs in order to enable automatic generation of high quality document metadata. This is the main focus of this research. To do this, this research focuses on exploring different approaches through Automatic Metadata Generation (AMG).

AMG provides methods for generating metadata without manual interaction using computer program(s) to interpret the document and possibly the document context. AMG is based on the observation that information that equals the desired metadata, directly or indirectly, may already be contained in the documents or in the context:

- **Visual descriptions**: By viewing the document through its native application or as a print-out, visual characteristics can be seen, such as the paper format and promotion of specific sections (e.g. some text with larger letters).
- **Technical descriptions**: By analysing technical information from the document or the system in which the document is stored, other characteristics can be obtained such as: file size, file format and storage dates.

- **Intellectual content descriptions**: By analysing the user specified textual content of a document, the intellectual content created by the user can be determined, such as the actual letters used to stipulate the document title.
- **Context descriptions**: Documents are not published at random. There is a link between the document which is created and the place in which it is published. E.g. published site and publisher role at that site.

The author of the document has hence directly or indirectly specified the desired content of many metadata elements. This can be utilized as AMG strive to avoid excessive manual efforts when similar metadata can be generated automatically based on existing data sources [21, 36, 42, 43, 44, 62, 99].

The domain of digital educational documents, or "Learning Objects" (LOs), is especially vulnerable to false or missing metadata. This is important since it is vital for the users of such documents to retrieve the correct information for e.g. curriculum reading or research. Due to this need for detailed and educationally accurate LO descriptions, the international educational metadata schema standard IEEE LOM has been created [74]. This metadata schema standard is extensive, enabling a rich and detailed document description. The LO together with a file with IEEE LOM metadata are the basis for the document package format standard SCORM [2]. Storing both the LO and the metadata which describes the LO in a single package file enable easier distribution of LOs with rich metadata descriptions; A combination which should be of considerable value for all educational purposes, as it would enable sharing of relevant LOs in a manner which we do not see today. A vital clue to why we do not see more SCORM LOs or other usages of the IEEE LOM is the metadata schema complexity. This complexity makes creation of IEEE LOM metadata a skill which has to be taught and demands plenty of time in the creation of LO metadata. Neither of these issues are currently the mainstream: Few people have the required knowledge to create IEEE LOM metadata, and of these people only a handful have the time to describe LOs with such metadata and packing them into a SCORM package. If IEEE LOM metadata could be automatically generated with the sufficient metadata quality, this would enable sharing and retrieval of educational resources in a scale we do not see today.

The current situation is illustrated at the Norwegian University of Science and Technology (NTNU). Here a Learning Management System (LMS) is used by students and lecturers to publish thousands of LOs yearly. This vast archive of educational resources does not promote sharing of LOs. No educational metadata is created. Hence, the search engine for LO retrieval does not have a data foundation needed to enable efficient sharing of LOs. Hardly any LOs in this LMS are reused. As a result, much human resources are used to recreate similar LOs each time there is a need for the LO. This limits sharing of knowledge within the organization. It limits sharing of knowledge with third parties. And it limits research and discovery of new knowledge by not enabling to build upon existing knowledge.

All of these issues would have been addressed if educational metadata could have been automatically created, especially if the metadata were to follow the IEEE LOM as this would enable sharing of LOs on a global scale. Still, we do not see the presence of

computer programs that can achieve such a metadata creation task. At the best, we see computer programs that enable generation of a limited set of metadata from a specific document collection, typically keywords from English documents in the PDF file format. Such strict requirements are not practical at NTNU. The published LOs do not share such homogeneous characteristics. At the NTNU, there is an extensive range of subjects and educational levels taught. There are a number of languages in use and hardly any restrictions in terms of document templates and file formats that the lecturers and students need to use. The LMS is designed to allow sharing of LOs regardless of file format and file content. As a result, the publishers of LOs have an extensive freedom to express themselves. This freedom is a major challenge for AMG algorithms, as there are no strict guidelines which characterize all the published LOs. Though, if such a set of AMG algorithms could be developed, this would make them usable within any educational context not only at this University, but on a truly global scale. There is a vast need for AMG algorithms that can generate rich, high quality metadata descriptions from non-homogeneous documents.

Even with such publishing freedom as described above, most of the published LOs at the NTNU LMS are Microsoft (MS) Office-based documents, such as MS Word, MS PowerPoint and MS Excel. At the same time, hardly any of the AMG based research efforts currently conducted is based on such file formats. Search engines also show that they have considerable challenges in accurately describing such documents. There is a vast need for AMG algorithms that can generate rich, high quality metadata descriptions from document file types which are actually used, rather than having to base efforts on a converted document version with characteristics that differs from the original document.

1.2 Problem Outline

Document collections at home, at work and “everywhere in between” seem to be growing explosively. This is while the efforts of enabling efficient retrieval of the right documents seem to be standing still. The existing research efforts in locating virtually identically formatted documents within a limited subject area just do not cut it when faced with our real-world challenges. We need efficient document retrieval regardless of how the documents look. We need efficient document retrieval regardless of what subject the document is about.

Metadata has been used for centuries by archivists and librarians to describe key characteristics of documents, in order to enable efficient document retrieval. Now everyone needs metadata in order to enable efficient document retrieval. For all types of documents. In all languages. For all subjects. The AMG efforts need flexibility and logics. This research expresses how such flexibility can be achieved and how this framework can be used not only to generate vast amounts of entities spanning a range of elements, but also how to achieve the high metadata quality essential for practical use of metadata for retrieval purposes.

When we have the desired metadata, we can exploit usages of metadata. One type of documents that are seldom shared with metadata descriptions are so-called Learning Objects; Documents intended for knowledge sharing by combining a document with technical and educational metadata of how and when the document is intended to be

used. Currently sharing of LOs is very limited. One major reason for this is the extensive complexity of the metadata schema and the high knowledge requirements metadata registration places on the author. This thesis will show how much of the required educational metadata can be automatically generated and packaged into a LO along with the document with a minimum of human efforts and limited user know-how requirements. By enabling this, this thesis' efforts could vastly increase the practical usefulness of LOs and enable sharing of digital knowledge regardless of geography.

1.3 Research Context

This research was initialized with the title "Digital Library and Learning". This thesis was soon guided towards metadata and the wonderful opportunities that arise for sharing knowledge when describing documents with rich metadata description. This brought us to the various metadata schemas which have been created for describing educational resources with general and educational metadata. Here this thesis faced its first challenge: We have simple and more complex educational metadata schemas, but hardly any documents are shared using such metadata. Often the datasets dedicated to a project or schema consisted of only a few handfuls of documents where the "largest" datasets were in the range of a few hundred documents. That is nothing compared to the millions if not billions of resources present on the Internet. So we have documents and we have educational metadata standards. Why aren't these standards used to enable efficient sharing of educational documents?

This thread brought this thesis towards the topic of automatically generating metadata, and the need for a framework to scientifically determine the quality of metadata entities. This has become the cornerstone for this thesis.

This thesis was to a large extent conducted as an individual task with guidance from the supervisors, financed by NTNU for four years. After this period efforts were conducted voluntarily.

This thesis has been inspired by other AMG-related projects and by search engines. These projects have shown possibilities for AMG, but also how narrow their field of view is, restricting their usefulness to perform when not paired up with exactly the correct documents. As for the search engines, most are privately founded and regard their AMG efforts as a trade secret. However, this thesis has been able to evaluate their results. And in the eyes of this thesis, these results were not up to par. These results have since been documented using the framework for determining the quality of the generated entities.

There were no datasets available for this thesis to use that contained diverse documents. As a consequence, this thesis contacted all teachers at NTNU in order to grant this thesis access to their courses' shared documents. This thesis is ever grateful to all the teachers who granted this thesis access, all of whom are listed in the Acknowledgement chapter.

A second dataset was retrieved from an Auditing firm in order to compare the quality of automatically generated metadata. In addition, the auditing firm was used to illustrate how to use document templates to promote desired usage.

Due to this being mainly a one-man research project, the human resources were limited. This thesis has chosen to focus on State-of-the-Art analysis and development of methodology for generating high quality metadata and LOs. This thesis' efforts needed to be limited in terms of development of executable program code. Programming of actual search engines is also outside of scope for this thesis.

This research is focused on documents that are actually being shared. And as the dataset showed, most of these authors distribute documents in MS Office or PDF file formats. So why not explore the metadata of "other" file types with potentially more "exotic" flavours? Such as MPEG7, MPEG21 files, or OpenOffice (LibreOffice) files. Well, because in this dataset such files were not shared. This thesis can document that OpenOffice were not used to create *any* of the published MS Office files. Regarding the published PDF-files, there are possibilities of these being based on Latex or OpenOffice. Sadly, the converting process over to the PDF file format is not lossless and as the study has shown, the metadata included in PDF files are strongly polluted by false or questionable entities.

1.4 Research Questions

The goal of this research is to:

> *RQ1: Find methods to automatically generate metadata from non-homogeneous document collections for promotion of educational resources.*

To do this, an analysis of the actual document file content, the so-called "document code", is central to learn about the content of each document. Basing AMG efforts around the document code can enable detailed, structured and correct metadata from non-homogeneous documents. To achieve the research goal, the following questions are answered:

RQ1.1: What is the quality of automatically generated document content (embedded metadata and document formatting)?

RQ1.2: Can AMG approaches be combined or selectively used on a document-by-document basis?

RQ1.3: Can AMG enable automatic generation of complex sets of metadata, enabling usage of advanced Learning Object document formats, such as SCORM?

1.5 Research Objectives

This research explores the following objectives:

RO1: Examine how commonly used content creation software (applications) use document code to store metadata, formatting data and intellectual content.

RO2: Document the kinds of metadata, formatting data and intellectual content that are contained in the document code of commonly used document formats.

RO3: Substantiate how document conversion between incompatible document formats influences the metadata, formatting data and intellectual content of the resulting document code.

RO4: Explore the possibilities for metadata extraction based on the document code and the consequences these efforts have on the quality of the generated metadata.

RO5: Explore the possibility of using the document code in combination with or directly as the data source for other extraction efforts based on visual characteristics and natural language AMG technologies, without the need for content presentation applications.

RO6: Explore the possibilities for using AMG technologies to assist in generation of advanced and complex to create resources, such as LOs in the SCORM format.

1.6 Research Design

This research needed to base its efforts on diverse documents in order to experience the effects of different document creation user environments and to gain documents with diverse visual and intellectual content. These documents were analysed in regards to their document contents and in regards to generation of metadata. The results of these analyses' were evaluated using an existing framework for measuring "quality".

The environment in which the document is created and maintained greatly affects the resulting documents. When you know what you are looking for it is often visible if the user that has driven the document creation process, or a system enforce environment control has been executing when the document was created and maintained. The commercial LMS called "It's learning" [81], which is used by NTNU, has been used for this project. Such systems are also known as "e-learning" systems. "It's learning" offers a system controlled environment where system-specific document types can be created and where stand-alone documents (documents created outside of a system controlled environment) can be uploaded by lecturers and students. This system provides access to documents created in a system controlled environment, some with content validation, along with uploaded, original and converted, stand-alone documents. The documents were published from courses in a multitude of subjects, including medicine, informatics, education and fine art.

An analysis was performed to document the characteristics of documents created in the system controlled environment. Such characteristics cannot be determined on this stage based on the stand-alone documents created and maintained in a user controlled environment.

A quantitative analysis of about 4000 LOs was performed to analyse the embedded metadata found in the retrieved documents. This was done in order to determine the availability and correctness of the embedded metadata. The quantitative analysis was concentrated on elements that could be partly or entirely judged as valid or false. This

analysis revealed that virtually no documents shared on NTNU's LMS contained either an educational metadata description or an informative description. So few documents had been given a semantic description besides the "Title" element, that it is highly questionable if the documents authors and publishers were aware that such content could be stored as part of the document. This research found it evident that a number of entities stored as part of the document were not created by the user. This includes technical elements such as file format and a number of time and dates. But it also includes elements like "Title" and "Creator" with entities with little or no resemblance to the title and creator name presented when viewing a print-out of the document. This research found considerable uncertainty regarding the quality of the gathered document metadata and regarding the awareness to metadata by document authors and publishers.

Ninety-one percent of the stand-alone documents uploaded to the LMS were in PDF, Word or PowerPoint document formats. The qualitative analyses of stand-alone documents consequently concentrated on these file formats. This analysis was performed to explore the possibilities for extracting metadata based on the document code. These extraction efforts were based on elements from the document that the embedded metadata did not reflect, or when no embedded metadata were present. By converting MS Word and MS PowerPoint documents into their Open XML formats, this research was able to gain full insight into the document code. This research therefore used the Open XML formats to analyse the range of possibilities available. These efforts were undertaken to explore the possibility of using alternative AMG approaches, where the document code was not suitable for generating the desired elements. The qualitative efforts were focused on the generation of the following elements: "Characters", "Words", "Pages", "Slides", "Creator", "Title" and "General. Language" (the language of the documents' intellectual content).

The Open XML format was chosen as the case document format over the competing document format OOXML, because MS provides full functionality to convert from binary Word (DOC) and PowerPoint (PPT) document formats to Open XML. Such functionality is not available for OOXML. Using OOXML would require the use of a third party converter application, which would introduce increased uncertainty regarding the interpretation of the existing, proprietary document format and the use of the new OOXML format. Using the MS converter also avoids the risk of contaminating or changing the document's content when it is converted.

The research results were evaluated using a framework for measuring "quality" presented by Lindland et al. [96]. This framework categorizes "quality" based on (1) Syntax, (2) Semantics and (3) Pragmatics. Additionally, supplemental quality terms were used based on Bruce et al. [17]. This framework supplements Lindland et al. [96] by including dedicated metadata quality terms for completeness, accuracy and provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility.

The international educational metadata schema standard IEEE LOM [74] was used to generate a common vocabulary and to define the content of specific elements and their

valid value spaces. However, this research is not restricted to this specific schema, and hence covers elements and aspects that are not included in this standard.

If the metadata descriptions and the LOs were distributed in a single LO-specific package, this would enable sharing of LOs with metadata descriptions. Many file formats have this option of including metadata descriptions, though rich description such as that is enabled by IEEE LOM is fare from common. Though, it is possible to include an IEEE LOM description to any educational document by creating a document package based on the SCORM standard specification [2]. This thesis presents how SCORM packages can be automatically generated with IEEE LOM metadata descriptions also automatically generated.

Topics from these research efforts have been sectioned into smaller subjects and presented as publications on various scientific conferences.

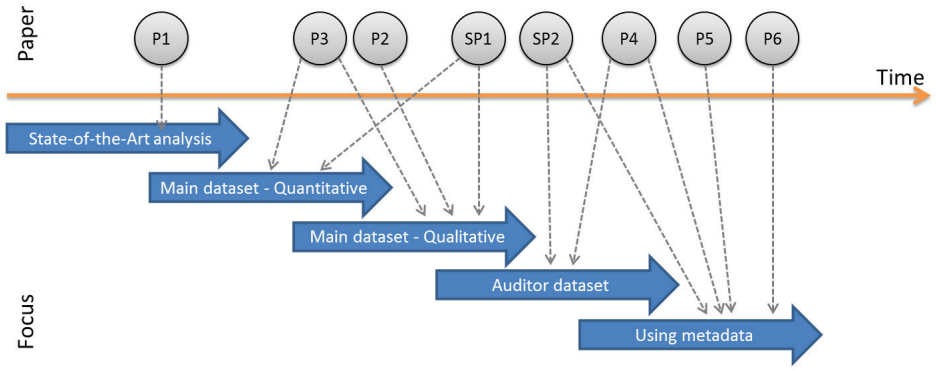


Figure 1: Timeline of main research focus and papers over time

Figure 1 illustrates the main research focus and papers over time. As it shows, this research has been through various phases, firstly to sketch over existing AMG efforts and their strength and weaknesses, secondly to retrieve a large dataset from NTNU LMS and analyse it, third to perform more in-depth analysis of selected documents and topics, fourth practical usage of AMG and the generated metadata, and fifth analysis of the second dataset from an Auditing firm and how to use document templates to promote desired usage.

1.7 Papers

This chapter gives a short introduction and presents the relevance of published papers and secondary papers. The primary, published papers have been published at respected international conferences. The secondary papers (SP) have either been published at NTNU or have yet to be published.

P1 Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg, *"Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment"*, Proc.of WEBIST 2007, March 3-6, 2007, ISBN 978-972-8865-77-1, pp. 427-432, Springer

Relevance to this thesis: This article introduces the challenges of generating metadata required for efficient retrieval and re-usage of resources on the Internet and on Intranets. The usage of Embedded metadata is presented as a topic, though there are quality concerns regarding these metadata. There is a need for other means of generating metadata without making this a burden on the end users.

P2 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Automatically generating high quality metadata by analyzing the document code of common file types"*, Proc. of JCDL 2009, June 15-19, 2009, ACM

Relevance to this thesis: The Document Code can be used to retrieve user specified data from a document and use these data as metadata. This opens for extraction of metadata across diverse visual document characteristics. However, there are a lot of data in the Document Code that is informative content suitable as metadata. This paper explores data sources and what data content that can be trusted.

P3 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Using the structural content of documents to automatically generate quality metadata"*, Proc. of Webist 2009, March 23-26, pp. 354-363, 2009, ISBN: 978-989-8111-83-8, ACM

Relevance to this thesis: The majority of documents published at the NTNU LMS are of Word, PowerPoint or PDF-file formats. This paper seeks to verify the quality of Embedded metadata and Extractable metadata. This paper introduces the concept of using the Document Code to combine AMG efforts in order to achieve higher quality metadata results.

The dataset retrieved for this paper were also used to write a report regarding usage of the LMS at NTNU called It's learning [88].

P4 Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg, *"Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing?"*, Proc. of IEEE DEST, IEEE Computer Society 2010, pp. 206-221, 2010, ISBN 978-1-4244-5553-9.

Relevance to this thesis: The Webist 2009-article used a visually and subject vice extremely diverse document collection. However, all document collections do not share these characteristic. This article focuses on a document collection

that should be strictly and precisely formatted: Conference papers. This article explores why high quality metadata are still not being generated.

P5 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects"*. Proc. of IEEE DEST 2009, June 1-3, INSPEC, 2009, ISBN: 978-1-4244-2345-3, 10.1109/DEST.2009.5276729

Relevance to this thesis: We now know that AMG algorithms can generate high quality metadata from visually diverse documents. In this article we go one step further and explore how such metadata can contribute to sharing of educational resources. In order to do this, the automatically generated metadata is used to generate a SCORM Learning Object containing a resource usable for educational purposes, and a rich metadata description of the resource. Though, there remain challenges in terms of low metadata quality of selected metadata elements.

P6 Ingeborg Torvik Sølvberg and Lars Fredrik Høimyr Edvardsen, *"Creating Metadata is a Costly Manual Process – And it can be Automated"*. In: Antony Jose (ed.) "Digital Libraries and Knowledge Organizations." Macmillan Publishers India Ltd., pp. 356-362, 2012. ISBN 978-935-059-076-8.

Relevance to this thesis: The number of authors of digital documents is ever-increasing. Most of these authors do not have any relationship to metadata. The amount of digital documents which each and one of us have created has also increased extensively. The amount of digital documents which this results in will only continue to grow in the future. The combination of an increased number of authors, increased number of documents and limited knowledge of metadata should promote an increased need for AMG in order to enable efficient document retrieval. Still, the research efforts on AMG for document retrieval seem to be decreasing. This article presents a re-cap of why we should be focusing efforts on AMG.

SP1 Line Kolås, Lars Fredrik Høimyr Edvardsen and Leif Martin Hokstad, *"Use of It's learning at NTNU – a quantitative and qualitative study"*. Original title in Norwegian: "Bruk av It's learning ved NTNU – en kvantitative og kvalitativ studie". Internal stand-alone study report at NTNU, conducted by the Program for Learning with Information and Communication Technology (Program for Læring med IKT (LIKT)) in order to review usage of It's learning at NTNU. pp. 1-157. January 2008. Published at and by NTNU.

Relevance to this thesis: This rapport analyses the usage of the LMS It's learning at NTNU. It reflects upon how the LMS is used in the various courses and faculties at NTNU. It illustrates large differences in usage between the different faculties. It also shows how the LMS is used. Importantly for this thesis, this rapport shows how Learning Objects are described with metadata when being published, or rather how *extremely limited* the metadata descriptions

are when the metadata descriptions have to be registered manually. This shows the need for including AMG efforts to automatically generate metadata, so that human creation of metadata can be kept at a minimum.

SP2 Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg, *"Using Document Code to automatically generate high quality metadata: An Auditing case study"*, Not published.

Relevance to this thesis: This article validates the results of using the State-of-the-Art AMG algorithms by using these on a document collection with vastly different characteristics than in the NTNU LMS and in Conference papers. This article presents how inclusion of non-visual Meta tags in the document templates can vastly increase the AMG algorithm's ability to locate and retrieve user specified content of a particular type. Still, the obtained metadata quality is not perfect.

1.8 Contributions

This thesis has had the privilege to contribute with contributions including:

C1: Establishing an overview of research literature, projects and products using AMG and the quality of their generated metadata.

This thesis has conducted an extensive State-of-the-Art analysis of literature, projects and products that use AMG. These efforts have been combined with a framework for determining the quality of the generated entities to analyse the strengths and weaknesses of the various AMG efforts.

C2: Establishing that AMG efforts can be combined in order to expand the range of elements and entities that can be generated, but also to increase the quality of generated entities.

A major limitation of most AMG efforts is that they generate entities regardless of the data source. Hence, many AMG efforts generate low quality metadata due to usage of a low quality data source or usage of a less optimal AMG algorithm.

This thesis has shown how the Document Code can be used to gain direct access to the authors' contributed contents to a document. This can increase the quality of the generated entities vastly. This while not generating entities when other data sources or other AMG efforts can be used to generate higher quality entities.

C3: Establishing that AMG efforts can generate high quality metadata from non-homogeneous document collections, vastly expanding the practical usefulness of AMG.

Nearly all research into AMG is done with nearly identically looking and formatted documents. The usefulness of the generated AMG efforts is severely compromised, as the developed algorithms could have limited, if any, usefulness when used on a different document set. In a real word practical scenario documents seldom share so many visual characteristics. Many authors base their efforts on a blank document template. In companies corporate templates are commonly promoted. In academics the various publishing sites and conferences use their own templates. However, it is up to the authors to comply with the specified templates. And quite often there are major differences between intended usage and practical usage. These issues severely lower the quality of the metadata traditionally created by AMG efforts.

This thesis wanted to show that AMG could be used to generate high quality and rich metadata descriptions to all documents, regardless of their visual characteristics. The developed framework for AMG has achieved this goal by generating high quality and rich metadata descriptions to all document types due to (1) selection of the best data source, (2) selection of the best AMG algorithm and (3) quality assortment and re-execution of AMG-efforts if needed. This thesis has demonstrated the high quality metadata that can be generated from large collections of poorly formatted documents. This thesis has also demonstrated how the quality of the generated entities and the range of desired entities can be vastly expanded by using the document template to promote a specific usage of the document template. By working with the document template, document sections can be re-located from any document regardless of language of the intellectual contents[1] and visual characteristics. This thesis' AMG algorithms hence generate high quality metadata from all document types regardless of contents. The algorithms could hence have usefulness in many contexts, not just with a dedicated dataset. Though, this thesis has documented that if common characteristics are known of the dataset, these characteristic can be exploited to increase the data quality and possibly the range of generated entities.

C4: Establishing that AMG efforts can contribute extensively in promoting sharing of knowledge with the creation of sharable SCORM LOs containing the educational resources themselves and extensive metadata descriptions to enable efficient location and use.

This thesis has demonstrated that AMG efforts to generate high quality and rich metadata descriptions can be generated for educational metadata as well. This includes

[1] This thesis has documented successful AMG efforts on documents, even AMG efforts on individual document sections, in multiple languages, including various English languages, Norwegian, New Norwegian, Danish, Swedish, German, Greek, French and Spanish.

descriptions of intended use, targeted user group and skill level in addition to other technical and descriptive metadata. This thesis has demonstrated how the generated metadata could be formatted in accordance with simple as well as highly complex metadata schemas including schemas specially developed for describing educational resources. This thesis has documented usage of such automatically generated metadata combined with the original document in order to create shareable SCORM Learning Objects containing the educational resource itself and extensive metadata descriptions to enable efficient location and use. By using such efforts, the skill level required for creating SCORM LOs could be lowered extensively, while allowing more LOs with educational metadata to be shared.

All the research questions, Contributions and Papers are closely interrelated. Hence, all the research questions, Contributions and Papers contribute to each other in some way or another. The major relationship between the Research Questions, Contributions and Papers are presented in Table 1.

Table 1: Relations among Research Questions, Contributions and Papers

Research Question	Contribution	Papers
RQ1	C1, C2, C3, C4	P1, P2, P3, P4, P5, P6, SP1, SP2
RQ1.1	C1	P1, P2
RQ1.2	C2, C3	P2, P3, P4, P6, SP2
RQ1.3	C4	P1, P3, P5

1.9 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 presents State of the Art. The main focus here is on clarifying what terms such as "metadata", "metadata schema", "quality", "Automatic Metadata Generation", "Learning Object" and "Learning Object System". This is done in order to establish a common view of current research and other aspects which affect the scope of this thesis.

Chapter 3 presents more of the contextual background of this thesis, including motivation for why these research questions were selected.

Results from this research include a number of publications presented in Appendix A and B. However, the publications can only scratch the surface of the research which has been performed. Chapter 4 is therefore used to present a detailed view of the research, including descriptions of how documents and LOs are commonly created and quantitative and qualitative analysis which have been the basis for the publications.

Chapter 5 is dedicated to conclusions and reflections based on this research. This includes:

- A main conclusion.
- A summary of major contributions within the research field.
- A comparison between the objectives set in the beginning of this research, and the actual results of the research.
- Recommendations reflect upon experiences gained though this research.
- The conclusion is ended by a presentation of recommended areas for future research work efforts.

Chapter 6 presents exclusively references.

Appendix A presents the published articles. The topics of these articles include: Challenges of generating metadata required for efficient retrieval and re-usage of resources on the Internet and on Intranets, how the Document Code can be used to retrieve user specified data from a document and use these data as metadata, addressing quality issues of Embedded metadata and Extractable metadata from documents published at the NTNU LMS, AMG-efforts based on strictly and loosely formatted documents and automatic generation of learning objects that include rich educational metadata.

Appendix B presents Secondary Papers. This section consists of the abstract from a report about usage of It's Learning at NTNU, published at NTNU, and of a paper focusing on AMG efforts on documents retrieved from an Auditing firm. The second paper (SP2) was never published.

2. State of the Art

Chapter 2.1 presents definitions for "metadata" and "metadata schema" concepts, while Chapter 2.2 presents definitions for "quality" and "quality metadata." Chapter 2.3 presents the AMG concept, followed by a presentation of the different AMG methods that have been developed, and of projects that use each specific type of algorithm as their main AMG approach. This chapter presents in detail how the document code can be used to generate metadata. Chapter 2.3.5 presents previously described projects and systems in more detail in terms of their efforts and reasons for using these technologies.

2.1 Defining "Metadata"

The handling of information in organizations has become a vital day-to-day challenge, as more and more information is archived in vast computer systems in the form of digital documents. Digital documents can be based on individual, stand-alone documents, such as Adobe PDF, MS Word and MS PowerPoint documents created and maintenance outside of a system controlled environment, or be of system specific document types. Usage of digital documents has introduced many new sharing and efficiency opportunities for disseminating knowledge. However, the use of such systems can easily limit information sharing if the "correct" documents are difficult to locate. With a rapidly growing collection of documents, locating the correct document becomes ever more challenging.

Metadata can be used to give each document a description that includes the key properties of the document. These descriptions can be a part of the data foundation used for document querying and retrieval efforts by allowing new users to find out about the documents' existence and their most central characteristics. The commonly used and simple definition of metadata is "data about data" [13, 69, 111, 124]. This is not an informative definition, however. Therefore, a number of more informative definitions have been developed [55]. This research bases its efforts on one such detailed and informative definition:

> *Metadata, or structured data about data, improves discovery of and access to such information. The effective use of metadata among applications, however, requires common conventions about semantics, syntax, and structure. Individual resource description communities define the semantics, or meaning, of metadata that address their particular needs. Syntax, the systematic arrangement of data elements for machine-processing, facilitates the exchange and use of metadata among multiple applications. Structure can be thought of as a formal constraint on the syntax for the consistent representation of semantics.*
>
> Miller [104]

These metadata are based on a pre-determined and standardized metadata schema that presents possible description types (elements) and the valid content of these elements, called entities. The metadata descriptions can be a part of the data foundation used for document querying and retrieval by presenting the document and its most central

characteristics in query results. The creation of metadata descriptions is a major challenge because of high user knowledge requirements, timely metadata registration processes, high human costs and the on-going challenge of the publication of ever more documents. These issues can be reduced or even avoided entirely by enabling computer software to generate metadata instead of, or as a supplement to, manual metadata actions. Such technologies are known as Automatic Metadata Generation (AMG).

The collection of the metadata elements that describe a document is known as a metadata *element set* [111]. These element sets are commonly stored as a *metadata record*. A metadata record is commonly defined as "A syntactically correct representation of the descriptive information (metadata) for an information document" [69]. A metadata record consists of a set of attributes, or elements, necessary to describe the document in question [69]. Metadata records can be embedded as part of the document or stored in an external metadata record collection. Metadata records are frequently presented as the digital equivalent of the traditional library card created for library cataloguing systems. The syntactically correct representation of elements and entities is defined by a *metadata schema*. The metadata schema is a systematic and orderly combination of elements used to specify valid element types, the entities that they can contain, and how these element types can be used [69]. The metadata schema is therefore a collection of syntax, definitions and a presentation of the permitted value spaces.

Rodriguez et al expresses a concern regarding the quality of document metadata as more and more people contribute with shared documents on the internet and other shared networks and communities [120]. Rodriguez et al. propose usage of an algorithm for "inheriting" metadata from other documents with similar characteristics [120]. Similar research was performed by Naaman et al. for labelling photographs taken in a series shortly after each other [106]. Rodriguez et al.'s efforts [120] were based on identifying sections from documents of a similarly formatted bibliographical dataset, characteristics such as citations, author, organization and keywords were inherited between documents. This project received mean correctness rate results of less than 20 percentages on average. Given the advantage of similarly structured documents in the dataset (see article P3, p. 208), this results indicate a need for more research before quality metadata is achievable on text-based documents. However, on non-textual, multi-media based objects, the research of Naaman et al. shows promising results for heritage between objects [106].

The FAsTA project presents how manually created meta-tags on the internet (Folksonomies) can be used as data source for automatically generated document metadata [7]. Though, this paper does not explore the quality of the potentially generated metadata. Bateman et al. also studied usage of manually created meta-tags on the internet, though found it questionable if such meta-tags would we helpful to students even though experts had provided the meta-tags [12].

2.2 Defining "Quality" and "Metadata Quality"

Defining "Quality" is subjective. Many frameworks for defining quality have been developed that focus on different aspects of quality and the understanding of the

described resource. This research bases its efforts on the framework presented by Lindland et al. [96]. This framework categorizes "quality" into three category types:

- **Syntax:** "Relates the model to the modeling language by describing relations among language constructs without considering their meaning."
- **Semantics:** "Relates the model to the domain by considering not only syntax, but also relations among statements and their meaning."
- **Pragmatics:** "Relates the model to audience participation by considering not only syntax and semantics, but also how the audience (anyone involved in modeling) will interpret them."

In terms of this thesis, these quality categories relate to the following issues:

- **Syntax:** Analysis of the document formatting to see that it complies with the document's format standard.
- **Semantics:** Analysis of the entities presented to see if they are valid and in accordance with the document format's relevant metadata schema.
- **Pragmatics:** Analysis to determine if the user-interpreted properties are reflected in the metadata.

Lindland et al. presented validation of syntax quality based on (a) prevention: exclusion of unwanted content, (b) detection: finding faulty entities that are used and (c) error correction: replacing faulty entities with correct entities [96]. Syntactic quality is determined based on compliance with the given document's compliance with the format specification, along with compliance with the value spaces associated with the document format.

Semantic quality is measured based on two goals: validity and completeness. Validity relates to the schema definition for the valid entity of each element. The validity of the semantic content relates to whether or not the element presents an entity that is relevant to the document at hand. Completeness relates to the extent to which everything that can be said about an element or a collection of elements has been presented in the resulting metadata records.

Bruce et al. presented a more detailed definition of "metadata quality" [17]. Their framework focused on user expectations and less on technical aspects. They categorized "metadata quality" into seven categories:

- **Completeness:** Completeness reflects two issues: (1) The use of as many elements as possible; and (2) that the user's desired elements are present in the metadata records.
- **Accuracy:** The entities should describe the document correctly and factually.
- **Provenance:** There should be a record of who created the metadata.
- **Conformance to expectations:** Assumes that the users' expected elements are available.

- **Logical consistency and coherence:** Logical consistency relates to compliance with the local metadata schema. Coherence relates to whether the elements are made available.
- **Timeliness:** Timeliness relates to two issues: (1) Currency: when the document changes while the metadata remain unchanged. (2) Lag: when the document is disseminated (distributed) before some or all metadata is knowable or available.
- **Accessibility:** That the metadata are available to users and understandable to users.

This research does not have a focus on specific end-user services, but rather on the opportunities that exist for generating a data foundation upon which end-user services can be built. The actual usability aspects of metadata are therefore not a subject for this thesis, and user accessibility and conformance to expectations are also outside of the scope of this research. Timeliness based on lag relates to when metadata are created and is an issue for manual metadata creation efforts. AMG algorithms can be executed as part of the document creation or publishing process, which means that timeliness related to lag is not relevant to this research.

Accuracy, provenance, logical consistency and coherence and timeliness based on currency are relevant to this research. These categories are the same as presented by Lindland et al [96], although with additional clarification. This research uses an extended vocabulary to increase the accuracy of quality based analysis.

This research has its main focus on syntax and semantic quality. This includes analysis of document formats and the entities of most restricted value spaces. However, because the evaluation of selected elements' entities is closely related to the visual presentation of the document, pragmatic quality issues are evaluated for these elements. This relates to semantic elements, where there are visible properties against which comparisons can be made, and to the distinction between the number of logical number and technical number of document pages.

The quality scale is measured subjectively as:

- **Very high:** The dataset can confirm a high degree of correctness.
- **High:** The dataset can confirm a high degree of correctness, although more than a few exceptions were discovered.
- **Undeterminable:** The dataset could not verify either correct or false entities for the given element, so that a conclusion could not be drawn.
- **Low:** Systematic false entities were verified to be present.
- **Very low:** An extensive number of false entities were verified as present in the dataset.

2.3 Automatic Metadata Generation

There are two main methods for creating metadata: Manual creation and automatic generation. Manual metadata creation can be difficult to enforce due to high knowledge and time requirements. Since this is the current default practice, only a fraction of potentially available documents are described with learning object metadata.

AMG algorithms are sets of rules that enable access to data source(s), identification of desired content, and collection of these data and storage of the data in accordance with metadata schema. AMG algorithms can use the document itself and the context surrounding the document as data sources. This thereby allows the re-use of content that is already available, although the data is subsequently structured in accordance with the intended metadata schema. Collecting embedded metadata is known as metadata *harvesting* [62, 116]. The process by which AMG algorithms create metadata that has previously not existed is known as metadata *extraction* [63, 66, 122][2]. AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that does not reflect the document. This places considerable demands on AMG harvesting and extraction algorithms to guarantee that they use available data sources in desired ways.

The following chapters present the main concepts behind the different AMG methods and data sources currently in use, along with a description of their main strengths and weaknesses. Most projects and systems use a combination of AMG methods and data sources for generating metadata. For each method section, a selection of projects or systems is presented that use each specific method as their main AMG method and data source. Chapter 2.3.5 goes into more detail for each project or system to present the methods used, what metadata were generated, the conditions under which tasks were performed and other contributions made by each work.

2.3.1 Data sources for Automatic Metadata Generation

There are two main data sources that can be analysed for the harvest or extraction of metadata; these are document-based data sources, and context-based data sources, as shown in Figure 2 (p. 21). The literature contains a number of alternative terms for "document-based," such as "object-based" [99] and "document content" [21]. Meire et al. [99] also used the term "context-based", while Cardinaels et al. did not present a term for "context-based" [21]. Instead, Cardinaels et al. described the context's three main data types directly [21]. This research uses the phrase name "document", which in turn results in the source types: "document context," "document usage" and "composite document structure."

AMG can be used to create metadata descriptions based on the document itself, by performing document content analysis. The document-based content consists of all content found in the document code, or the technical, document format or system based formatting, the intellectual content created by the user(s), and the embedded metadata stored as part of the document code.

[2] There is currently inconsistent usage of these method names in the published literature: Here "extraction" is sometimes referred to as harvesting of existing metadata. This research will be using the definition given above; That this should be regarded as "harvesting," not "extraction."

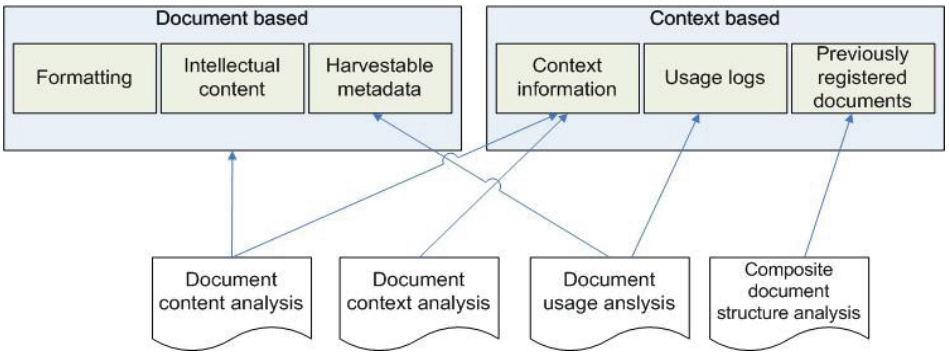


Figure 2: Data sources and related AMG analysis approaches

AMG can also be based on context information. The context information is commonly collected from the document publishing system where the document was published. However, other external data sources can also be used, such as document description databases that are outside of the domain of the publishing system. Context-based content can be divided into three main categories based on its content type: Document usage descriptions (Document usage analysis); metadata collected from prior versions of the document (Composite document structure analysis); and other information provided by the document context (Document context analysis). Selected types of document content analysis algorithms use context information to generate document metadata that is influenced by the context, such as document classification. Some document formats include usage information stored as document metadata, which can be used as part of the dataset used by document usage analysis algorithms. See Figure 2 for an illustration of the document formats and their relationships to AMG algorithm types.

The following Chapters present information on Document content analysis, Document context analysis, Document usage and Composite document structure and projects using these data sources.

Document content analysis

This first approach is based on analysis of the document itself. Document content analysis is the traditional method of performing AMG by examining the document itself to create the metadata [60]. There are two approaches to document content analysis: the collection of existing, embedded metadata from the document, which is called harvesting; or executing algorithms to create metadata from data sources that is not based on metadata, called extraction. Each of these methods has its own strengths and weaknesses, which are presented in detail in Chapter 2.3.2.

Selected types of document content analysis algorithms use context information to generate document metadata that is influenced by the context, such as document

classification, and the generation of document keywords, which is based on predefined, subject-specific keywords located in a context outside of the document.

Document content analysis efforts commonly combine harvesting and extraction, where each metadata element is based on specific data sources that are combined with specific harvesting and extraction efforts. Figure 17 (p. 45) shows an example of a document content analysis effort that harvests the "Author" element and extracts the "Title" element based on visual characteristics.

Document content analysis was performed by all the projects and systems listed in Chapter 2.3.5: AMeGA [16, 49, 56, 57, 60, 64, 71, 82], The Jorum project [84, 85, 91, 94, 97, 98], MAGIC [92], Metadata Analyzer [125], Metadataminer Pro [121, 122, 126, 134, 135].

Document context analysis

The second approach is analysis of the environment surrounding the document. Document context analysis based AMG methods collects data from the user's local environment for the creation of metadata. Such methods take advantage of the user being logged into a publishing service, such as a LMS, and the specific section of the publishing service where the publication took place.

Log-in information can be used to identify the publisher and the role that he or she plays in the context of the publishing sections that are accessed. This user profile can be used to generate a vCard consisting of possibly extensive information regarding the user and the user's role in the specified context. A vCard is a standardized and structured collection of user related information, including the person's name [80].

The context information regarding the specific section where the publication took place can describe a document on a more abstract level. These are descriptions that may relate to a collection of documents, such as all documents published in relation to a course, but not the individual document.

By applying differing levels of abstraction, increasingly document-specific descriptions can be introduced as the level moves closer to the actual document. These more specific descriptions are generated by including more elements, or by applying more appropriate entities from each abstraction level. As such, a tree of abstract levels can take the form like that shown in Figure 3. Here the entities set at Level 1 are transferred to the underlying level, Level 2, and so on.

The use of context descriptions offers special potential for educational documents, particularly because educational metadata are seldom retrievable from the documents themselves. This result in a need for alternative ways of generating metadata based on data sources other than the document content, while limiting the need for human interaction to generate the metadata. Context-based defaults can be used to specify default entities for a number of elements that reflect common metadata schemas, such as the IEEE LOM.

Level 1: Default values for the entire LMS.
Level 2: Default values for the specific University
Level 3: Default values for the specific Faculty at the University
Level 4: Default values for the specific Institute at the Faculty
Level 5: Default values for the specific department at the Institute
Level 6: Default values for the specific course run by the department
Level 7: Default values for the specific sub-section of
the course LMS area (e.g. folders)

Figure 3: Increasingly specific levels of context data

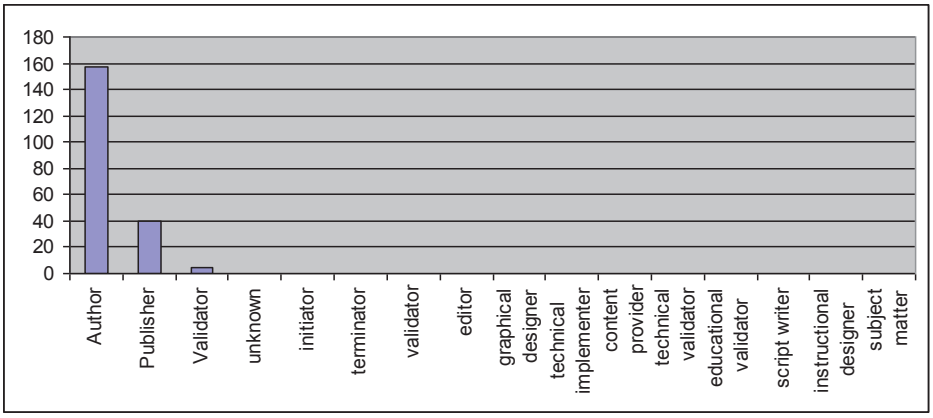


Figure 4: Use of the value space for "Lifecycle. Contribute. Role" based on Friesen [52]

Statistics gathered by Friesen presented how some specific entities from the available value space for specific elements are more commonly used than others [52]. These statistical data were collected from existing LOs published though different publishing services. Figure 4 shows one such restricted value space where close to 80% of the registered entities were of one specific type. This specific element presents the role of the contributing person for the specific document. By using such content as default entities, human efforts can be shifted from generating identical metadata to correcting the entities where the default is not correct.

A combination of levelling of content and entities based on statistical data can be used to generate a range of context descriptive metadata. Table 2 shows an example of such a dataset. Here the default entities were set at different levels. Each level presents its own set of default entities, and describes the educational context in more and more detail. The example also illustrates the replacement of a higher level entity with another entity: The language of the document (visible or audible) when the document is opened is

corrected from being Norwegian at the University level to being (British) English, which is used as the official language for that specific course.

Table 2: Using local context to create global metadata

	Existing metadata	IEEE LOM metadata
A)	LMS default	2.2 Status = Final 3.3 Metadata schema = LOMv1.0
B)	LMS University context	1.3 Language = NO 5.6 Context = Higher education 5.7 Typical age range = 18-
C)	LMS course context	1.3 Language = en-GB 5.5 Intended end user role = Learner 5.8 Difficulty = Very difficult 5.11 Language = NO
D)	LMS course sub-section	5.2 Learning document type = Exercise 5.9 Typical learning time = Pd7 [3]

No project has been found that uses this approach as its main AMG method. Projects using this approach as part of their efforts have been found for projects in the approach employed by Duval et al. (see Chapter 2.3.5 subchapter 1.1), the Jorum project (see Chapter 2.3.5 subchapter 1.4) and efforts by Jenkins et al. (see Chapter 2.3.5 subchapter 1.5).

In addition, context information is actively used as the data source for specific types of content extraction based metadata generators. In projects using this approach, the context information usually includes an extensive context description consisting of keywords, thesauri or ontology. This approach is therefore closely related to both the rule-based natural language approach and the machine learning method based on the natural language approach.

The major disadvantage of using document context analysis is that the metadata generated are not based on the document itself. Therefore, the metadata generated do not necessarily reflect the described document and hence may be incorrect as metadata for the specific document.

In addition, the different levels of default values need to be actively used to enable distinctions to be made between documents. As the number of elements that contain entities set at a high abstraction level increases, the likelihood that elements will differ among the metadata from the documents decreases. If these entities are incorrect, then the value of the metadata records is reduced. It is therefore important to correct

[3] This is coded information representing the entity “7 days.”

elements that contain entities that do not reflect documents that are representative of the specified abstraction level.

An analysis performed as a part of this research has determined that a document's heritage can also be shared among documents on the same "level"; for example, the document type "Note" is commonly used to describe all documents within the specific folder. Common usage includes creation of a description that explains how the user should use the documents in the folder. This data source can be used as an educational description. Similarly, document descriptions are assigned to the folder containing the documents, rather than the individual documents themselves.

When there is a relationship between documents, this relationship can be documented in the metadata descriptions, as when there is some relationship between documents in the same folder. Metadata relations can then be created between the documents in the folder. Similarly, all documents published in association with a course can be assigned relationships on a more general level.

Predefined relationships between documents based on explicitly created references are even more specific, such as "is part of." Then the targeting document's metadata can be automatically updated with the opposite document's relationships (here: "has part"). Direct references also include hypertext links, such as "references," and conversely, "is referenced by." More advanced ways of detecting relationships include: (1) pattern-matching with predefined topics within a subject, (2) manually located links based on pattern-matching, and (3) where automatic links are placed between pattern-matching and manual links [9].

Document usage analysis

This third approach is based on retrieval of information of actual document usage in order to generate metadata. Some elements reflect document properties where the intended usage pattern differs from the actual usage. In regards to the IEEE LOM schema, this is reflected in the element "5.9. Typical learning time." However, there can be extensive differences between intention and practice.

Computer systems can track and log actions performed by the user. These data logs can record the document's actual user group, the actual typical learning time, and so forth, instead of the intended user group and learning time. For example, for educational video-based documents, the typical usage time can be set by harvesting or extracting the video's play time. However, if the video is paused or stopped during the video presentation, then the usage time of the video will differ from the content-generated entity. For other document types, such as research papers, presenting an accurate entity for describing such properties can be difficult to set and can therefore end up not being used. In an educational environment where the differences between intentions and reality can be substantial, accurate document usage metadata can provide a valuable information source.

This research has not found related work that takes advantage of this data source. This is closely related to the fact that generating such a data source can be regarded as user

surveillance. For some element types, this entails recording the user's actions, such as the identity of the person who accesses each document, and the role this person has. Other elements require permanent surveillance of the user throughout the learning process. Data that relate to this are most likely being collected by the LMS and other publishing systems, such as data on document access and the use of content packages that are exclusively used in the specific user environment. This last example is found in SCORM content packages used with an LMS. However, the collection and use of these data are generally undertaken for administrative purposes, not to enable enrichment of the document metadata. Using these data to generate metadata may encounter political and moral obstacles.

Composite document structure analysis

By using Document context analysis described previously default entities commonly specified in the publishing system are given to documents. The fourth and final data source for automatically metadata generating efforts is based on heritage of metadata from related documents rather than other environmental sources. The use of a composite document structure has the potential of transferring metadata from existing documents to new documents. By doing this, efforts to generate metadata from prior versions of a document can be transferred to new versions.

The composite document structure approach can also be used to establish a heritage of metadata from individual documents to larger documents containing multiple documents. In such cases, the larger document can accept entities collected from all sub-documents it contains. An example of this is a document that contains an entire course, for which metadata are inherited from the lecture and exercise documents that made up the course document. This is also relevant for content packages, such as SCORM. Here the sub-components can generate metadata that describe the resulting content package.

Using a composite document structure enables the reuse of existing metadata. However, if no metadata are available, then this approach does not enable automatic generation of metadata. This approach is therefore more frequently used in formal document repositories than in LMSs, such as Digital Libraries or digital corporate archives. Digital libraries or digital corporate libraries are mainly where the major initiative for this AMG approach is presently found. One example of this approach is where the ECHO project designed a digital library service for historical films owned by large national audio-visual archives [30, 39]. They adapted the IFLA-FRBR model to this task by creating 4 main levels ("Work," "Expression," "Manifestation" and "Item") to increase the efficiency of generating metadata for new versions or sub-documents of existing documents [76].

Ochoa et al. express the need for using default entities and the heritage of entities from documents in an educational context [114].

Content of a document code

Current AMG efforts are based on one or more the previous four previously presented data sources. Hence, a fifth data source has been given very limited attention, commonly not even mentioned in AMG theory. That is, until now.

A document provides information not only from what the author presents, but also in *how* the presentation is executed. This is a form of information which is stored in the document file, though is not necessarily visually distinguishable from other document contents. Regardless of visible characteristics, if the data is part of the document, it must be present in the document file discoverable using document code analysis.

A "source code" is defined in computer science as any sequence of statements and/or declarations that are written in some human-readable computer programming language. Stand-alone document formats such as plain text (TXT), HTML and XML are sequences of statements and/or declarations that can be human-readable. These document formats are not computer programming languages, and hence do not comply with the definition of a source code. The main objective behind most document formats is not to obtain human readability, but rather to enable application usability. For example, the Open XML document format is XML-based, which allows human review. However, because of the potential complexity of XML-code, it is human readable only to a very limited degree. However, it does provide readability for the text-based content and the formatting of the sections where the content is located. Most current document formats are binary, relying on dedicated applications to interpret the document content before its intellectual content can be presented in a human-readable form, such as Word and PDF documents. Some document formats can contain applications, such as Word documents. The different properties of documents make the boundary between documents based on source code and not on source code blurry. Table 3 shows different properties of selected document formats.

Table 3: Common document formats

Format name	Format extension	Binary document format	Standardized document format	Embedded metadata	Usage metadata	Text-based intellectual content	Formatted template content
MS Word	DOC	Yes	No	Yes	Yes	Yes	Yes
MS PowerPoint	PPT, PPS	Yes	No	Yes	Yes	Yes	Yes
MS Excel	XLS	Yes	No	Yes	Yes	Yes	Yes
Adobe PDF	PDF	Yes	Yes	Yes	No	Yes	No
HTML	HTM, HTML	No	Yes	Yes	Yes	Yes	Yes
Text	TXT	No	Yes	No	No	Yes	No
Open XML	DOCX, PPTX, PPSX, XLSX	No	Yes	Yes	Yes	Yes	Yes
JPEG	JPG, JPEG	Yes	Yes	Yes	No	No	No

This research uses the term “document code” to refer to all content of a document, similar to what is meant by referring to “source code” for applications. This is done in order to avoid classifying different document formats as source code, some source code, and not source code. The content of documents created with LMS is also referred to as document code.

The document code consists of all the documents' stored data, based on the user actions performed, the template that was the basis for the document, and all data stored by the content creation software. In terms of stand-alone documents, the document code consists of all content within the document (file). System specific documents consist of all content that is present in the system’s definition of the smallest document type.

The document code of documents has traditionally been binary. Different document formats have used different binary coding. Gaining access to the document content has therefore required that the coding of the specific document format be understood. This has been further complicated by proprietary document formats for which binary coding is regarded as a company secret and is hence not fully revealed. This is true for commonly used stand-alone document formats such as MS Word, PowerPoint and Excel. Gaining access to the document content of such documents can therefore be very challenging. Figure 5 shows how some of the content of a Word document can be accessed, although all formatting and sectioning has been lost. Use of document code for binary document formats for AMG purposes has not been found by this research. However, commercial applications have been developed to harvest embedded metadata from stand-alone documents.

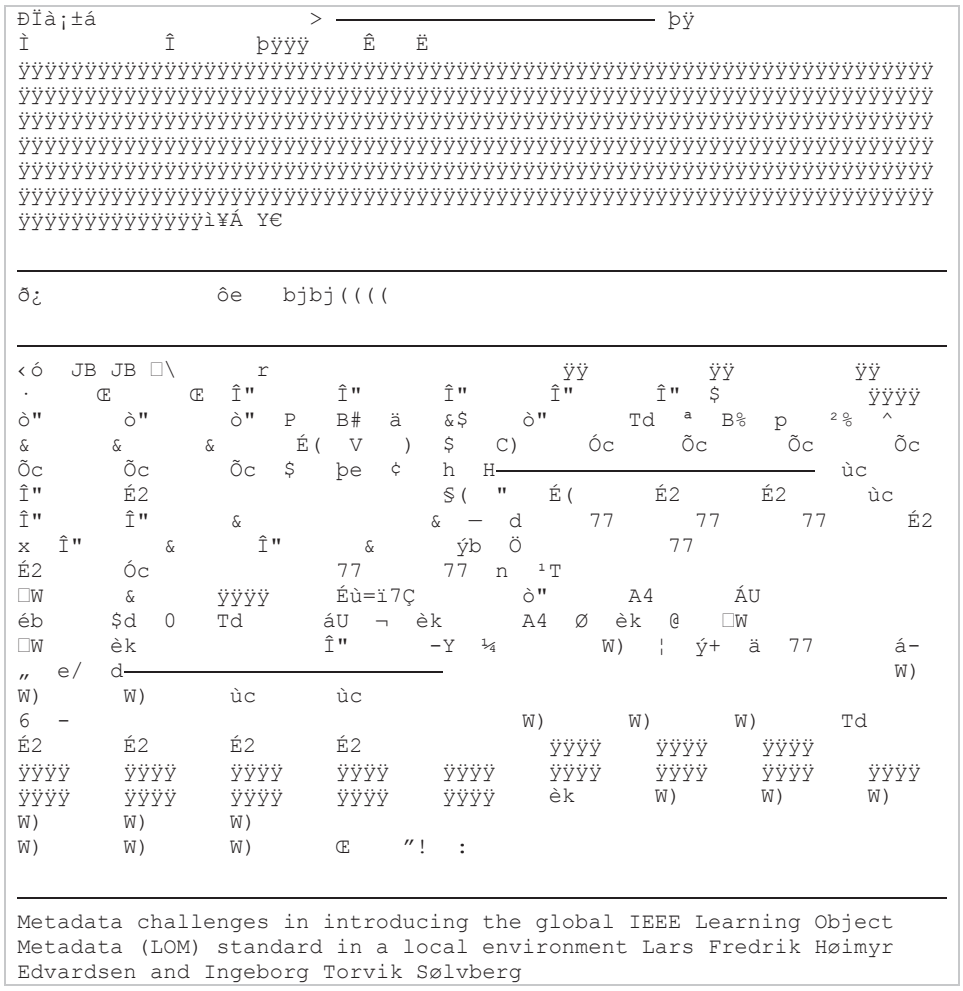


Figure 5: Binary document code for a Word document interpreted as text

```
%PDF-1.5
%µµµµ
1 0 obj
<</Type/Catalog/Pages 2 0 R/Lang(en-US) /StructTreeRoot 43 0
R/MarkInfo<</Marked true>>>>
endobj
3 0 obj
<</Type/Page/Parent 2 0 R/Resources<</Font<</F1 5 0
R>>/ProcSet[/PDF/Text/ImageB/ImageC/ImageI] >>/MediaBox[ 0 0 595.38
841.98] /Contents 4 0
R/Group<</Type/Group/S/Transparency/CS/DeviceRGB>>/Tabs/S/StructParen
ts 0>>
endobj
...
...
...
42 0 obj
<</Title(Metadata challenges in introducing the global IEEE Learning
Object Metadata \(LOM\) standard in a local environment)/Author(þÿ L
a r s   E d v a r d s e n   a n d   I n g e b o r g   S ø l v b e r
g)/Subject(Informatics, Webist 2007)/Keywords(IEEE LOM, Learning
Object Metadata, LOM, Learning Object, LO, Learning Management
System, LMS, metadata mapping, crosswalk, metadata
challenges)/Creator(þÿ M i c r o s o f t ®   O f f i c e   W o r d
2 0 0 7   \( B e t a \))/CreationDate(D:20070114151844)
/ModDate(D:20070114151844) /Producer(þÿ M i c r o s o f t ®   O f f i
c e   W o r d   2 0 0 7   \( B e t a \))>>
endobj
51 0 obj
<</Type/ObjStm/N 321/First 2874/Filter/FlateDecode/Length 4726>>
stream
```

Figure 6: Document code for a PDF document

Easier access to embedded metadata has been enabled by a number of binary document formats, such as later versions of Adobe PDF, JPEG and MP3. Here the document code is split into two logical sections: A text-based embedded metadata section and binary code of the remaining of the document, the document body. This enables access to the document metadata without the need to understand the remaining document content. Figure 6 shows the same document as in Figure 5, but converted into PDF to enable easier access to the metadata.

Open source document formats have come, or are about to come into common public use. This enables increased interoperability and reusability of documents. Here the entire document format has been made available and is possibly standardized. This enables third party applications to understand the entire document's content. Open source document formats are also more easily read by humans, which makes it easier for humans to create applications for these formats. Such openness has allowed for many AMG projects that use a plain text HTML document format to harvest metadata, including the projects AMeGA [60], LOMGen [123] and by Xue et al. [134]. Plain text HTML has also been used to analyse the document to locate references to other

documents, as done by Jenkins et al. [82]. Figure 7 shows the HTML header ("<head>"), which includes embedded metadata elements from the header sub-tags. Open source document formats have also been introduced to system specific documents, and are the standard for many online document publishing systems, e.g. Blackboard [15] and It's learning [81].

```
<html>
<head>
  <meta http-equiv=Content-Type content="text/html;
      charset=windows-1252">
  <meta name=Generator content="Microsoft Word 12 (filtered)">
  <title>Metadata challenges in introducing the global IEEE Learning
      Object Metadata (LOM)standard in a local environment</title>
</head>

<body lang=EN-US link=blue vlink=purple>

<table class=MsoNormalTable border=0 cellspacing=0 cellpadding=0>
  <tr>
    <td><p class=MsoTitle><span lang=EN-GB>Metadata challenges in
       introducing the global IEEE Learning Object Metadata (LOM)
       standard in a local environment</span></p></td>
  </tr>
  <tr>
    <td><p class=Author align=center style='text-align:center'><span
       lang=NO-BOK>Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik
       Sølvberg </span></p></td>
  </tr>
```

Figure 7: Document code for an HTML document

The document formats for the MS Office content creator software suite are about to be changed to document formats based on Open XML or possibly OOXML. At present, the use of Open XML has been included in the MS Office 2007 suite, including a lossless conversion application for converting binary MS Office document formats into Open XML. In the context of this research, there is full backwards compatibility with earlier MS Office document formats. Hence, older documents can be re-saved in an Open XML document format while retaining their original formatting, and without contaminating the data sources used in this analysis. Examples of Open XML document code can be seen in Figure 16, Figure 60 and Figure 66.

Open XML documents are zip-compressed archives. They contain multiple text-based XML files, folders and other objects included in the document. There are dedicated XML files for document properties, including the embedded metadata elements, the main document, header, footer, slides and spread sheets. A special file (the "core.xml"-file) contains the embedded metadata elements, and uses the Dublin Core metadata schema standard. Additional document format specific properties and metadata are

available from the "app.xml"-file. This includes the Word document's "Title" element and headings, and PowerPoint slide titles. Each document format has a dedicated folder where the main documents are stored.

2.3.2 Approaches for Automatic Metadata Generation

The previous chapters described data sources that can be used to automatically generate metadata. This next chapter presents methods for using these data from these sources and turning these data into informative metadata.

AMG algorithms are constructed to take advantage of one or more available data sources. The algorithms are constructed based on rules that enable them to gain access to the data source, identify desired content, and collect and store this information in accordance with a metadata schema. These rules are executed when the AMG algorithm is initiated. If these rules are manually created, they are referred to as "rule based algorithms," or if they are created by an application, they are called "machine learning algorithms." AMG algorithms that use existing, embedded metadata are referred to as "harvesting" algorithms, while algorithms that create new metadata, are referred to as "extraction" algorithms.

Document content analysis is the main approach used to generate metadata from previously unpublished documents. These algorithms base their efforts directly on the document code of the document or use a content presentation application to present the desired document content before AMG efforts are undertaken. Current document content analysis efforts are based on four different approaches:

1. **Harvesting of embedded metadata.** This approach uses embedded metadata created by the document creator software or by the user and stored as part of the document [14, 57, 63, 121, 135]. These metadata are placed in a specific location of the document, enabling harvesting algorithms to locate and harvest the metadata without a need for interpreting of the content of the document. See Figure 9 and Figure 10 for a dataset example and Figure 8 for an illustration of the four different content analysis approaches. This approach is vulnerable to generating false metadata if the data sources do not contain high quality metadata.
2. **Extraction based on visual appearance.** This approach uses a special content presentation application to generate a visual representation of the document before executing rules to extract content that is based on the visual appearance of the document [49, 56, 85, 49, 91, 97]. The content presentation applications commonly present the documents as if presented in the documents' native content creation software or as a print-out. The visual representation is used as data source for rules adapted to identify and extract specific visible document content. See Figure 11 for a dataset example. This approach is vulnerable to generating false metadata if the documents do not share the visible appearance(s) for which the algorithm has been developed. Hence, such algorithms only perform as desired on pre-determined document types.

3. **Extraction of metadata based on natural language.** This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed [16, 60, 82, 92, 94, 98]. Such algorithms commonly include collection of unique words and comparisons of the document vocabulary against reference ontology for generating keywords, descriptions and subject classification. See Figure 12 for a dataset example. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages or document sections in different languages.
4. **Extraction based on the document code.** This approach uses the document code directly, without the need for additional content presentation applications to interpret the document content. This enables full and direct access to the entire document's content. This includes template identification, template content identification and formatting characteristics, regardless of visual characteristics and the language of the intellectual content. See Figure 13 for a dataset example. This approach requires that the extraction algorithm be able to interpret the content of the document. This can be a challenge due to binary document formats, proprietary, not standardized document formats and otherwise complicated document formats. Current, popular document formats are binary (PDF) or non-standardized (Word & PowerPoint). This has limited the research based on document codes to HTML documents [82]. With the emergence of new document formats, this thesis will explore the use of the document code from Word and PowerPoint document formats.

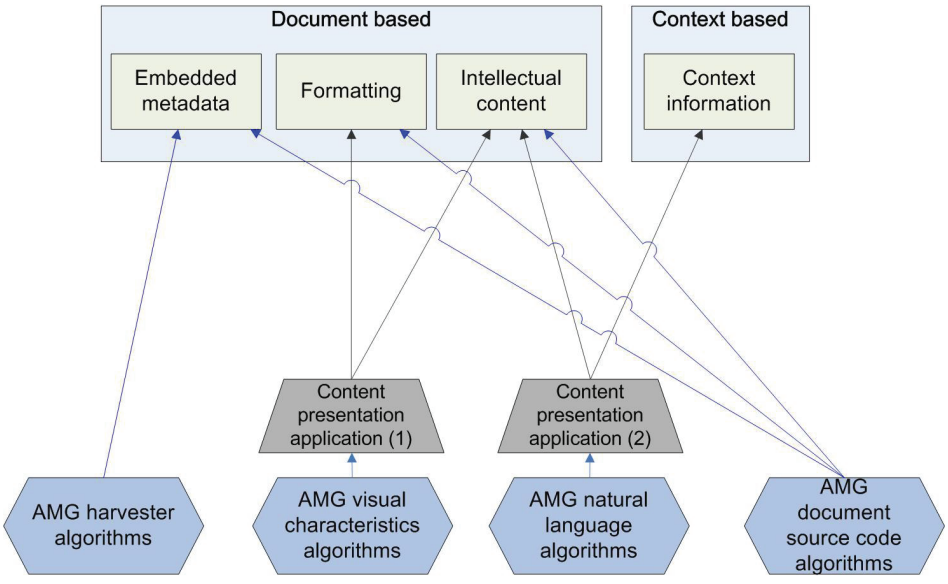


Figure 8: AMG content analysis algorithms and their data sources

```
<head>
  <meta http-equiv=Content-Type content="text/html;
      charset=windows-1252">
  <meta name=Generator content="Microsoft Word 12 (filtered)">
  <title>Metadata challenges in introducing the global IEEE Learning
      Object Metadata (LOM) standard in a local environment</title>
</head>
```

Figure 9: Embedded metadata from a Word document stored as HTML

```
<o:DocumentProperties>
  <o:Title>Metadata challenges in introducing the global IEEE Learning
       Object Metadata (LOM) standard in a local environment</o:Title>
  <o:Subject>Informatics, Webist 2007</o:Subject>
  <o:Author>Lars Edvardsen and Ingeborg Torvik Sølvberg</o:Author>
  <o:Keywords>IEEE LOM, Learning Object Metadata, LOM </o:Keywords>
  <o:Description>The world of closed LMSs ...</o:Description>
  <o:LastAuthor>Lars</o:LastAuthor>
  <o:LastPrinted>2007-01-12T13:59:00Z</o:LastPrinted>
  <o:Created>2007-01-08T10:02:00Z</o:Created>
  <o:LastSaved>2007-01-24T10:22:00Z</o:LastSaved>
  <o:Pages>6</o:Pages>
  <o:Words>3534</o:Words>
  <o:Characters>20147</o:Characters>
  <o:Lines>167</o:Lines>
</o:DocumentProperties>
```

Figure 10: Embedded metadata from a Word document stored as XML

Figure 11: Visual characteristics of document content

METADATA CHALLENGES IN INTRODUCING THE GLOBAL IEEE LEARNING OBJECT METADATA (LOM) STANDARD IN A LOCAL ENVIRONMENT

Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg

Dept. of Computer and Information System, Norwegian University of Science &Technology,

Sem Sælands vei 7-9, NO-7491, Trondheim, Norway

{lars.edvardsen,ingeborg} @idi.ntnu.no

Keywords: IEEE LOM, Learning Object Metadata, LOM, Learning Object, LO, Learning Management System, LMS, metadata mapping, crosswalk, metadata challenges

Abstract: The world of closed Learning Management Systems (LMS) is being replaced by open systems for sharing and reusing digital Learning Objects (LOs) between users, courses, institutions and countries. This poses new challenges in describing these LOs with detailed and correct metadata. This information background is needed for querying services to perform accurate queries for LO retrieval. . In this paper we present metadata specific challenges when converting from a local LMS with proprietary metadata schema to a global metadata schema. We have uncovered extensive LO description possibilities based on the existing, local LMS, registered metadata, its LO types and the local context. Files can contain extensive metadata descriptions, though require special attention. We have confirmed that technologies developed as crosswalks are valid for usage in this projects for a one-time metadata transferral. However, transferring of all local metadata elements can result in incompatibility issues with other LMSs. This, even when keeping with the global metadata schema.

1 INTRODUCTION

The use of digital Learning Objects (LOs) such as slides, figures, exercises and exams are increasing on all educational levels…

Figure 12: Natural language of document content

Figure 13: Document code of document content

The following chapters present the following methods for generating document metadata: Harvesting, Extraction based on visual characteristics, Extraction based on natural language and Extraction based on the document code.

Harvesting of embedded metadata

The approach of harvesting existing, embedded document metadata can be regarded as the easiest way to generate document metadata. A number of commonly used document formats can contain embedded metadata as part of their document code. See Table 3 on p. 28 for examples of some document formats and their embedded metadata. These metadata can be created by content creation software, by users, or by both. There are two main reasons for including embedded metadata:

- To allow content creator software to correctly identify the document format and enable encoding and interpretation of the document content in the intended way. For example, there are currently eight versions of the Adobe PDF document format where distinction between versions is based on version metadata.
- To enable more usability for the document creator. These embedded metadata are therefore also commonly displayed in different user interfaces to enable the user to more easily locate the desired document. For example, the song name, album, release date and artist name are frequently displayed for MP3 sound documents.

Specific document formats can contain extensive embedded metadata descriptions, such as MS Office document formats, which include logistical metadata regarding the creation date, last saved date and last printed dates, semantic metadata with the name of the user who performed the previously listed actions, title, keywords, description and technical elements regarding the number of characters, words, pages and slides of which the document consists. JPEG images can contain an XML-based section (EXIF) that can contain data regarding the camera settings when a picture was taken, geographic location (GPS coordinates) and technical descriptions of the image (resolution (dpi), dimensions (horizontal and vertical number of pixels), etc.). Adobe PDF documents can contain multiple metadata sections, allowing metadata based on multiple metadata schemas to be included in a single document. An extensive range of elements is thus supported.

A selection of content creation software automatically generates embedded metadata for semantic metadata, such as the author name, title and keywords. However, there can be problems regarding the correctness of the entities that they generate. For example, the MS Office suite of applications and Adobe Distiller (which converts original documents into Adobe PDF documents) generate elements that do not always reflect the document at hand. This is due to the use of default entities, which are elements that are not updated and elements that are replaced when the document is converted to an alternative document format. This has made use of the embedded metadata challenging for AMG algorithms to generate metadata that reflects the specific document at hand. This research has not found content creation software and document formats that store meta-metadata. As a consequence, the author of specific metadata elements cannot be determined based on meta-metadata.

Different document formats have different approaches regarding where in the physical file the metadata are stored, how these data are coded and the metadata schema they use. Gaining access to the embedded metadata therefore requires knowledge of the structure and interpretation of the specific version of the document format. There is therefore no general method of gaining access to embedded metadata. Projects using embedded metadata are therefore concentrated on specific document formats. The most common document format studied is HTML, because it is open source, is frequently used on the Internet and uses a text-based document code format. This makes it easier to gain access to the metadata and other document content than working with binary document formats, such as PDF and Word.

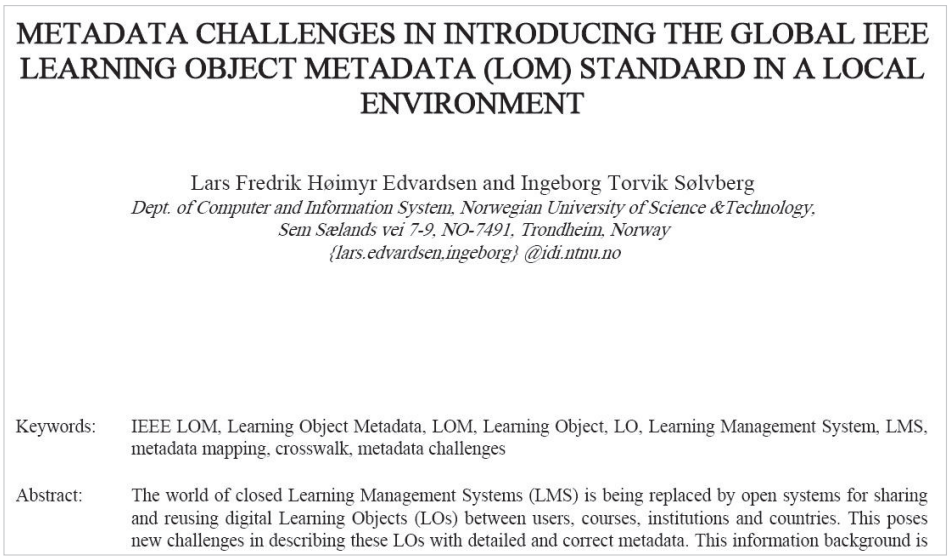
Initiatives that have used harvesting as their main AMG method include the Greenstone Digital Library [64] and the Jorum project [57, 84, 131, 135]. Special commercial applications have also been developed to harvest metadata from a range of different stand-alone document formats and their proprietary metadata schemas, such as Metadataminer Pro [126] and Metadata Analyzer [125].

Extraction based on visual characteristics

Metadata harvesting is limited to the specific elements that are present in the document. Content creation software (user applications) is known to systematically generate false metadata. This is a reason for why many document projects do not use this data source. As a consequence, many projects enforce extraction of metadata rather than harvesting.

This approach uses a content presentation application to generate a visual representation of the document. Such applications can attempt to present the document as if it were presented in its native content creation software or as a print-out. This representation is created based on the document formatting and the intellectual content created by the document user(s). The visual representation is used as the data source for rules adapted to identify and extract specific visible document content. See Figure 11 for a dataset example.

The algorithms based on visual characteristics use the visual appearance of the document to identify document content. The rules expressed in Figure 17 points (a) to (d) express visual conditions. The advantage of this approach is that rules can be created to identify multiple elements found in an individual visual document.



METADATA CHALLENGES IN INTRODUCING THE GLOBAL IEEE LEARNING OBJECT METADATA (LOM) STANDARD IN A LOCAL ENVIRONMENT

Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg
Dept. of Computer and Information System, Norwegian University of Science &Technology, Sem Sælands vei 7-9, NO-7491, Trondheim, Norway
{lars.edvardsen,ingeborg} @idi.ntnu.no

Keywords: IEEE LOM, Learning Object Metadata, LOM, Learning Object, LO, Learning Management System, LMS, metadata mapping, crosswalk, metadata challenges

Abstract: The world of closed Learning Management Systems (LMS) is being replaced by open systems for sharing and reusing digital Learning Objects (LOs) between users, courses, institutions and countries. This poses new challenges in describing these LOs with detailed and correct metadata. This information background is

Figure 14: Visual characteristics of a paper

The example from Figure 14 is based on a standard template for scientific papers based on the guidelines for the LNCS format provided by Springer [127]. By identifying the visual appearance of a document based on this template, rules can be used to identify and extract each content section; in this example, this applies to the title, author(s), affiliation, e-mail address, keywords and abstract. Other, more universal rules have been proposed for use, such as using the first line of text as the title or using the text string with the largest font as the title.

The use of rule-based content extraction based on visual characteristics as the main AMG method has been attempted by Flynn et al. [49], Liu et al. [97], Kawtrakul et al. [85], Li et al. [91] and by GESTALT [56].

The major hurdles for AMG algorithms for extraction using rule-based visual characteristics are their complexity, general validity and preciseness.

- Different documents with different visual presentations require their own rules. These algorithms are vulnerable to extracting unintended content from documents that have visual characteristics that differ from the documents of the dataset for which the algorithm rules have been developed.
- Identification of each document type can be a considerable challenge.
- This AMG approach relies on using a content presentation application to interpret the document content before the extraction efforts can be performed. The document presentation algorithms give their own perspectives of the document content upon which the continued analysis is based. This makes for a

data source that differs from the original document content. Non-standardized document formats and document formats that are intentionally interpreted in different ways by different applications are particularly vulnerable to inconsistencies between the actual content of the document and the content presentation application's presented content. This can be visualized by comparing Figure 11 (visual characteristics), Figure 12 (natural language) and the actual content of the document in Figure 13.

- The algorithms are vulnerable to collecting and analyzing content that is not part of the main document content, such as content from headers and footers.
- Different rules need to work efficiently together.
- There are issues regarding prioritizing of data sources and different rules.
- It is difficult to create a labyrinth of rules needed to successfully generate valid metadata entities for a range of document types.

There is therefore an extensive demand for human efforts to generate rules, determine rule weights and to adapt the rules to work together to generate the desired results. This is further complicated if the document formats are evolving, e.g. if a new content creation software version uses the document format in new ways. Then the AMG algorithms need to be updated to tackle documents created using both the old and new software. As a consequence, rules that were previously correct can become incorrect, or may require a re-shuffling of the labyrinth of rules to determine the best candidate entity.

The use of the natural language approach has its weakness in multi-lingual environments where documents can be of more than one language or may include document content sections that are in different languages.

Both the visual characteristics and the natural language approaches are additionally influenced by their reliance on content presentation applications that need to recreate the document in the dataset before any analysis can be performed. These applications distort the content of the document, as the content of the document code differs from the datasets generated. This can be visualized by comparing Figure 11 (visual characteristics), Figure 12 (natural language) and the actual content of the document in Figure 13. The document presentation algorithms give their own perspectives of the document content upon which the continued analysis is based.

Extraction based on natural language

Natural language rules have been developed as an alternative to rules based on visual characteristics. This approach also uses special content presentation applications to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed. Such algorithms commonly include the collection of unique words, and comparisons of the document vocabulary against a reference ontology for keyword generation placed in the document publishing system's context information. See Figure 12 for a dataset example. Natural language based algorithms can function by comparing content from different sections of the document against each other and by weighting the value of specific words and phrases.

The natural language approach requires extensive local knowledge to adapt the algorithms to the way local users employ their language and vocabulary. The algorithms need to handle different forms of words, synonym words and synonym phrases without confusing or mixing documents. To cope with this, technologies such as thesauri and ontology are frequently used. However, the generation and maintenance of these technologies is manually labour intensive. It requires extensive knowledge of how the language is used. This makes the vocabularies that are developed case- or subject-specific, which limits their general use. This limits the use of such technologies to the specific subjects and local contexts for which they were developed. This is therefore a solution that has been adapted to subject-specific document collections. The use of rules based on the natural language approach is most common in generating entities for more general elements, such as summaries, descriptions and keywords, although this method has also been used to generate titles.

Rule-based content extraction using the natural language approach has been used a main AMG method by AMeGA [60, 82, 94, 98] and MAGIC [16, 92].

A sub-division of the rule-based approach using natural language was developed using "folksonomies." Folksonomies practice collaborative tagging of documents, allowing multiple persons to create a reference and "tag" them with keywords. These are services usually aimed at general public use, and hence involve a user group consisting of other than professional metadata labellers. Since all the content is shared, folksonomies can be used to generate ontology based on the content specified by the community of users. This allows the use of more freely chosen keywords instead of a controlled vocabulary of traditional ontology [132]. Al-Khalifa et al. demonstrated the use of the folksonomy approach to generate IEEE LOM metadata [8]. The Melt project used folksonomies running under the European Schoolnet [37, 100]. These efforts are concentrated on semantic elements. The approach of using folksonomies is interesting, though still in an early development phase.

The major hurdles for AMG algorithms for extraction using rule-based natural language are their complexity, general validity and preciseness.

- The use of the natural language approach includes many sub-processes that increase the complexity of the developed extraction algorithms.
- This AMG approach relies on using a content presentation application to interpret the document content before the extraction efforts can be performed. The document presentation algorithms give their own perspectives of the document content upon which the continued analysis is based. This makes the data source that differs from the original document content. Non-standardized document formats and document formats that are intentionally interpreted in different ways by different applications, are particularly vulnerable to inconsistencies between the actual content of the document and the content presentation application's presented content.
- The data sources used for comparisons for documents that are based on another language may be inappropriate; in other words, if a document written in

Norwegian were to be analyzed and compared to an English dataset, then the Norwegian words would not be trunked, masked or stemmed correctly. Algorithms such as the one presented by [82] would present only Norwegian words instead of subject-specific, unique words. Hence the wrong words would be analyzed. Similar issues would crop up if other document content types were to be submitted, such as a document written about informatics in a system that uses a dataset vocabulary developed for the discipline of medicine. The natural language approach therefore needs to be language- and subject specific in order to generate the best results.

- The algorithms are vulnerable to collecting and using document content that is not part of the main document content ("document body"), for example, collecting content from headers and footers.
- The different rules need to work efficiently together.
- There are issues regarding the prioritizing of data sources and different rules.
- It is difficult to create a labyrinth of rules that are needed to successfully generate valid metadata entities for a range of document types.

There is therefore an extensive demand for human guidance in generating rules, determining weights and adapting the rules to work together in generating metadata. This is further complicated if the document formats and subjects are evolving. Then the AMG algorithms need to be updated to address both new and old challenges. As a consequence, rules that were correct earlier can become incorrect or require a re-shuffle of the labyrinth of rules to determine the best candidate entity. In addition it is becoming more common for documents to be generated in multi-lingual user environments, which further complicates the situation for natural language based algorithms.

Extraction based on the document code

Extraction based on the document code uses the document code directly, without the need for additional content presentation applications to interpret the document content to create a usable dataset for AMG efforts. This enables full access to all document code content without the potential contamination from content presentation applications as a result of their interpretation of the document code. Basing AMG efforts directly on the document code avoids many of the challenges that face extraction algorithms based the visual presentation of a document. Using the document's code allow the AMG algorithms to gain direct access to the user-specified document content. This avoids having to use technologies such as OCR or other conversion applications to gain access to the document content and its formatting. This is true regardless of the visual presentation and the language of the intellectual content used for the document and avoids:

- The need for judgment regarding the visual document content (such as font sizes and content placement). Instead, facts regarding the content can be used. Rules based on visual characteristics could hence be made more accurate.
- The unwanted analysis of data sources (such as headers and footers).

Additionally, the approach enables the collection of complete content sections, such as complete text boxes in PowerPoint documents, even when the text spans multiple lines.

Basing extraction efforts directly on the document code requires the ability to correctly interpret the content of the document format's document code. This has until the present been a major obstacle due to binary document formats, proprietary, not standardized document formats and otherwise complicated document formats. These additional challenges in gaining access to the document content and in the interpretation of this content have limited the number of projects that have based their AMG efforts on this approach. The exception to this is the projects that take advantage of the easy access to the text-based content of HTML document code to harvest Meta-tags for embedded metadata, to extract hyperlinks, and as data source for other rule-based or machine-learning-based AMG algorithms. However, a document's document code can be used to obtain much more detailed document descriptions. Due to the limited use of this approach, this research presents more details regarding this approach regarding opportunities and challenges.

The document code of commonly used stand-alone document formats such as Word, PowerPoint and PDF is enabled to contain extensive visual and non-visual formatting. Even HTML documents can provide extensive document descriptions based on the document code. The document code can contain information regarding the template upon which the document is based. Such template information is present in all MS Office documents and HTML documents created using MS Office applications. These facts enable AMG algorithms to be adapted to specific templates, and allow the AMG algorithms to perform more accurately because the document and its known template type are more closely related. This reduces the need to judge the type of document that the document actually is which in turn allows the correct identification and extraction of more elements if they are present in the template. For example, the general purpose "normal.dot" template for Word documents does not include any document descriptions that are usable in this context. However, other templates can provide extensive information regarding the document's visual appearance and content sections, such as organization, journal and usage adapted templates. One example is the NTNU template "e_brev.dot," which consists of the official department letter format, presented in Figure 15. Based on this template, specific content sections can be identified by analysing the document code's template-standard section names and their section content. In this example, this includes extensive university, faculty, department and author information, references, dates and the number of pages. Figure 16 displays how a specific selection of the document content can be identified and its content made available for rule-based extraction. Here the "Our reference" section is located (visually in the upper right corner of Figure 15), and the entity "REFNUMBER 1234" is displayed for possible extraction. Similar efforts can be undertaken with scientific document content, by identifying journal or conference templates and performing extraction in accordance with the template used. The document code can also be used to gain access to other content created by the user, such as references, illustrations, figures and tables, all without relying on visual characteristics.

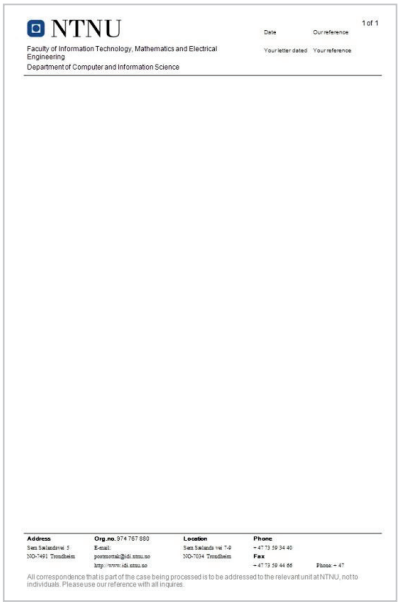


Figure 15: Official NTNU letter template in Word format

Figure 16: Open XML document code of Figure 15 once filled in

It is to be expected that other alternative AMG approaches could result in more desired results for specific elements. For example, algorithms developed based on the natural language approach can generate more representative keywords than the user-specified keywords available through the document code. The content presentation applications that are used by AMG algorithms based on visual characteristics and natural language can be adapted to present the document content in specific ways. These presentations can be better representatives of the document content than what is specified in the document code. Basing AMG efforts on the document code enables the extraction of content only if the specified content is available; for example, if no reference number were included in the example above, then no content would be available in the reference section of the document. Alternative means of generating metadata could then be executed.

The document code can be used to obtain document descriptions that are automatically generated, though not presented as metadata, as when the language used in the document is automatically included in MS Office documents to enable the use of spell-checkers. This can also be seen in Figure 16, where some sections are marked with “en-US” (English-US). The MS Office application practices automatic labelling of text-based content sections or even single words. It is therefore possible to distinguish between the specific content sections and the language used in that section. This would

be a valuable tool upon which to base natural language algorithms, since it can exclude content in languages not covered by the natural language algorithm. This can avoid one of the major challenges in introducing language-based algorithms in a multi-linguistic user environment, such as a university.

The potential of using this approach has been limited by the general understanding of document codes. This situation is currently changing, as commonly used document formats are being moved towards non-binary, standardized document formats based on XML code. Such formats have been introduced by Microsoft (MS) for their MS Office document formats (MS Word, MS PowerPoint and MS Excel). Their new document format is based on the Open XML standard. MS supplies a lossless converter application between the "old" binary and the "new" XML-based document formats. This enables full insight into the document code of these documents formats. This in turn allows for a range of AMG efforts based directly on all the content of the document code, but without document content distortion.

In contrast to efforts based on visual characteristics and natural language, extraction based directly on the document code does not have to result in an entity. If the extraction is undertaken of pre-specified content, then AMG algorithms that rely on document code may return no result if the desired content is not located. This allows for the efficient usage of alternative AMG algorithms in cases where the document code does not provide the desired result. For example, if the "Our references" section of Figure 15 is blank, the AMG algorithm should return a blank result. These section-specific properties reduce the need for judgment regarding the actual content of the entity obtained once a usage pattern has been determined.

The close reliance on the document code is also the greatest weakness of this AMG approach, however. If the document code semantics and formatting present something other than the desired content, then false content will be generated as metadata.

Additionally, the content of the document code may have been generated by multiple authors:

- The user may have created the content. If this content is in compliance with the intended document schema, then this is often desired content.
- The content may have been inherited form a template, such as an old document. The document metadata and formatting can therefore reflect another document than the one that has been analyzed. For example, several NTNU document templates contain elements with pre-defined entities, such as Creator = "O. Rakel" and Title = "Line one."
- One or more content creation applications may have been the author of metadata, semantics and formatting. MS Office documents can contain language formatting tags that are not used, as one example. Some applications can also be the author of content presented as part of the document's intellectual content, such as in converter applications that include visual commercial document content.

Commonly used document formats do not include meta-metadata that describes the author of the document content. Distinguishing between desired content created by the author and undesirable content created by another party will then have to be based on other reasoning approaches.

Content created by the user can be falsely formatted, using the existing document template formatting to promote other content than the schema-specified content, as when the user's name is included in the "Title" section and vice versa, which are then accessed as visible characteristics for these sections of the template in use. Analysis of the user's actual usage of the document therefore needs to be undertaken to ensure that the user's intentions are reflected in the metadata generated.

Efforts need to be enforced to obtain knowledge about the templates that are used, in order to avoid the use of false template-based content, such as collection of information that reflects other documents.

Use of the document code require extensive knowledge of how applications employ document formats in order to avoid data sources that present content that does not reflect the actual document. This includes new usage patterns that result from new applications or application versions.

2.3.3 Development of AMG rules

The previous chapters presented ways to automatically generate metadata using various data sources and approaches. However, generation of one specific metadata using one specific approach does not provide the rich and high quality metadata that this thesis strives for. To do so, the different data sources need to be exploited using the most suitable approach. Rules are such instructions that describe conditions in which the various data sources and approaches should be combined in order to generate the desired metadata. Such efforts can vary from simple one-source one-algorithm rules to complex multi-conditional rules based on artificial intelligence.

AMG algorithms are constructed based on rules that enable access to the data source and identification of desired content to collect this information and store it in accordance with metadata schema(s). These rules are executed when the AMG algorithm is executed. Such rules can be manually created, which are referred to as "rule-based algorithms," or they can be created by an application, commonly called "machine learning algorithms." The following chapters present the development of AMG rules based on rule-based algorithms, along with machine learning rules.

Rule-based algorithms

The manual creation of rules requires extensive existing knowledge of the documents at hand. The performance of these algorithms reflects the knowledge of the algorithm creators, their knowledge of the documents and their judgment regarding how their rules should be executed. This in turn results in the appropriate creation of rules, prioritizing of rules and selective use of rules. As a consequence, the rule-based approach is dependent upon having personnel who define the rules needed to achieve the desired results. This can require extensive local knowledge of how documents are used and how

the documents are presented, particularly if such efforts are based on visual characteristics or the natural language approach. This also requires redefining rules and their use as the dataset evolves. Figure 17 presents a set of rules that can be combined in generating metadata for the "Author" and "Title" elements. The metadata were generated by harvesting the "Author" element and extracting the "Title" element based on visual characteristics.

a) The author name is located in the document's metadata section that is identified as the "Creator" element in this section.
b) The title element is located on the first page of the document.
c) The title element uses the largest font on the page.
d) The title element is in 80% of documents written with bold letters.
e) The title element is in 20% of documents written with italic letters.
f) The title element shall not start the word "Draft."
g) The title element must start with letters, not symbols.

Figure 17: Example of rules for rule-based algorithms

The rules used in extractions can be absolute or be given a "weight." Absolute rules (or rules with a maximum or minimum weight) contain definitive requirements, excluding all content that does not conform to the set rule requirements. The points (a), (b) (e) and (f) of the example above are absolute. The rules can also be less strict, having a weight set to other values than the maximum or minimum. This allows for the retrieval of candidates before an evaluation takes place to select the most likely candidate entity as the metadata entity. Rules (c) and (d) are examples of this. If there are multiple candidate entities resulting from the algorithm, then these rules can be used to rank the candidate entities in order to select the best candidate entity.

Initiatives that have used rule-based algorithms as their main AMG method include: AMeGA [16, 49, 56, 57, 60], Greenstone Digital Library [64, 82], The Jorum project [84, 85, 91, 97, 98] and MAGIC [92, 121, 135].

Machine learning algorithms

Machine learning has been developed from the field of Artificial Intelligence (AI) to avoid the need for human judgment in the task of creating rules and determining the weights applied to each rule. These algorithms gather statistical data that is then used to optimize rules and weights to maximize the end results. Machine learning algorithms first reach their potential with large document sets, when sufficient statistical data has been gathered in order to form the most favourable rules and weights. The rate at which the Machine learning algorithms gain experience depends on the algorithms that are used. A range of alternative algorithms, or models, has been developed that reflect the use of specific properties, such as the Super Vector Machine and the Variable Hidden Markov Model.

Initiatives that have used machine learning as their main AMG method includes: Hu et al. [71], Liddy et al. [94], Seymore et al. [122], Xue et al. [134] and Yahoo [135].

Major challenges with rule-based approaches

Both manual rule-based algorithms and computer-generated machine learning algorithms have high knowledge requirements for initial implementation. Both approaches also face extensive challenges as their dataset evolves. This implies getting to "know" the "new" documents and the development of new rules and weights without lowering the correctness rate achieved with the initial and traditional document formats.

This is a process that requires the human rule developers to gain experience with the new data, while the machine learning approach requires that the necessary statistical data be collected and analysed before the approach can be adapted.

2.3.4 Conflict handling and trust

In order to automatically generate high quality metadata, there is a need for execution efforts besides the AMG algorithms themselves. The AMG algorithms might generate more entities than the metadata schema allows. And the generated entities might not be up to the desired quality requirements. Additional execution efforts are often needed to select among candidate entities and if the prime candidate element(s) should be used.

When there is the potential of obtaining more than one return for an entity, there is a need for a conflict handling function. Multiple, alterative entities have the potential to be present for all available data sources and between data sources. For example, the document that was the basis for Figure 15 and Figure 16 contained two "Title" elements, one from the existing, embedded metadata, and one from the document code retrieved through metadata extraction. Some metadata schemas allow multiple synonym elements, though such elements can make logical inconsistencies and hence lower the logical consistency and coherence quality of the metadata records.

Establishing which data source to prioritize can be challenging, because many documents do not present meta-metadata regarding who has created the conflicting entities, when this occurred or a description of how the data were gathered. It can therefore be challenging to determine if an entity was created by a user or by an application and if the entity reflects the latest or earlier versions of the document. Validation of the available data sources and analysis of how their entities are created is needed in order to establish guidelines for handling conflicts.

Meire et al. [99] proposed using manually operated "Conflict Handling Methods" to resolve such issues. The MAGIC system practices manual correction facilities after the AMG algorithms have finished executing [92]. Liddy et al. showed prioritizing of data sources when generating metadata [94]. The human generated entities were given the highest importance and hence were the most trusted. In a survey performed by Greenberg regarding the quality of metadata, she concluded:

"Results also indicate that extracting metadata from META tags created by humans can have a positive impact on automatic metadata generation"

Greenberg [62]

This is an issue of great interest, since few people are aware of the metadata being created. In the AMeGA project, 58.1% of the participating *professional metadata labellers* recognized that they had been using applications that automatically generated metadata [60]. That is a low number, considering that commonly used applications such as the MS Office package and Adobe Acrobat have been automatically creating metadata for many years, which include titles, author names, dates and statistical data. Most applications that generate audio and video (still image and moving image) automatically generate extensive metadata descriptions.

Figure 18 illustrates that metadata professionals believe that the metadata labelling process can be automated based on the DC schema. However, it also shows that there is a difference in trust regarding the entities that can be generated. The user groups therefore reported that they would like the ability to make manual corrections after the AMG processes were executed [60]. Such functionality would have to be adapted to the user group(s) at hand in order to satisfy local requirements and preferences.

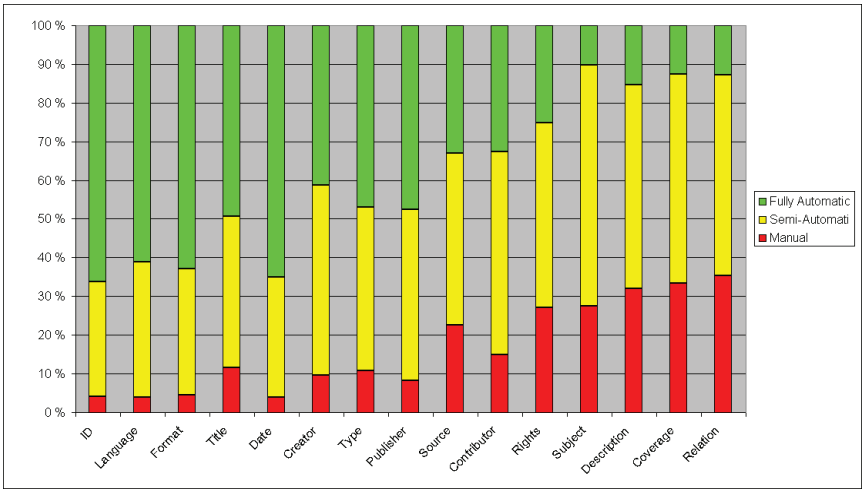


Figure 18: Degree of trust in AMG by metadata experts, from Greenberg et al. [60].

2.3.5 Projects and systems using AMG

The label "AMG" is not particularly much used in computer systems and services. Still, a large number of research projects, search engines and document retrieval services are

extensive users of AMG technology. This chapter explores a number of computer systems and services using AMG.

Firstly we will look at complete document sharing systems that exploit AMG. Secondly we take a look at search engines and how they perform. Thirdly, we look at various research systems and projects which have explored usage of rule-based AMG efforts. Finally we take a look at research systems and projects utilizing machine learning AMG efforts.

Development of the AMG system structure

In this first of four subchapters, we take a look at complete document sharing systems that exploit AMG. These are commonly computer systems based research efforts for storing documents, automatically generating metadata and storage of the generated metadata in accordance with a specified metadata schema.

Duval et al.

Duval et al. have been working to create the IEEE LOM standard, to ensure interoperability between educational schema and to explain and demonstrate how AMG can be used for generating IEEE LOM document descriptions. Their efforts are concentrated on creating a framework for AMG rather than specific algorithms. Cardinaels et al. presented how to create a simple-to-use AMG user interface and an internal, module-based program structure [20, 21]. They proposed using AMG methods for document content and context analysis before presenting the metadata to user for manual correction possibilities and conflict handling. This model was further explained and tested by Meire et al. [99]. They presented how this model can be used to generate metadata of multiple schemas. They concluded that a system like this, using the algorithms developed in previous papers, can generate metadata of equal quality to the metadata found in the ARIADNE document repository. They also stated that the most limiting factor for achieving quality metadata was using “not updated data,” or metadata that reflects a prior version of the document, and not the current version. This is results from many applications, including the MS Office applications, which do not update all their metadata each time the document is saved. All the metadata elements used in the in-depth analysis are vulnerable to containing data from prior document versions.

Related to this work, Ochoa et al. presented how closed LMSs with documents can be used to generate Learning Object Repositories (LOR) where the documents are freely available [114]. Their efforts are based on mapping content from an existing LMS with documents described according to standardized metadata (IEEE LOM or ARIADNE LOM). This group also presented a semi-automatic generation process, where the system generates metadata, and then presents them to a metadata author (person) who selects the elements that are desired for inclusion in the individual document. They proposed using default values for specific elements when no data source was found that could determine the “right” entity. Using default entities can be linked to the course section or other logical organizations for heritage of entities.

They concluded that such a framework can significantly reduce the amount of human documents needed to generate metadata. However, their approach requires humans to

decide and make corrections to false metadata. A higher degree of automation in the registration process would be desirable.

Duval et al. are involved with the MELT project, which is developing a system using folksonomies for describing Learning Documents [37, 100]. To do this, they are using existing tools developed by the Katholieke Universiteit Leuven/ARIADNE and the European Schoolnet. They have a goal of generating metadata that are language and culture independent. This project is on-going.

Greenstone Digital Library

The Greenstone Digital Library is a freeware digital library package that can be used as a basis for local digital library services [11, 64]. The main goal is to offer software that promotes digital library services. As part of this package, Greenstone has included tools for AMG that are intended for use as documents are placed within the DL. Greenstone converts all text-based documents into their Greenstone Archive format for storage. They use an extended version of Dublin Core as their metadata schema. Metadata are generated as a part of the converting process.

Greenstone uses the harvested modified date ("Moddate" or "Last Saved") as the "Date" element [63].

The original "Title" element is harvested and used as "Title" for PowerPoint, Word and HTML documents. The file name is harvested and used as "Title" element for Excel documents. The first line of text is used as the "Title" element for PDF documents. Metadata extraction is used for PDF documents. Their "Title" element is generated by collecting the first line of text. Greenstone applies a filter to remove the entities "Page 1" and "1" from the "Title" elements from PDF and Word documents. If no "Title" element is registered (after filtering), then the first 100 characters of the document body are used as the "Title."

Greenstone harvests the "Author" elements from PDF and MS Office documents before mapping them as "Creator" elements [63]. It also harvests the "Generator" element, which specifies the application that created the original document, although this element is only used in original HTML documents. The PDF document format uses the element name "Creator" for such data. MS Office documents use the element name "Application", but no mapping is provided between the "Creator" and "Generator" and between the "Application" and "Generator" elements.

For PostScript documents, the number of pages is harvested, although no number of pages or slides is registered from other document formats. No number of characters or words is registered as metadata, although the application uses these data sources for generating other metadata entities. Greenstone provides services for generating key phrases and automatic classification. A composite metadata structure is demonstrated by having collection level metadata that describe multiple documents. Sub-collection level metadata are also available for describing smaller selections of the document collection.

AMeGA

The Automatic Metadata Generation Applications (AMeGA) project analysed current AMG initiatives, conducted a survey among metadata experts regarding AMG, and created a report of recommended functionality for AMG applications [60].

This project presents a range of metadata elements that can be harvested from commonly used document formats. They also present projects that are working on metadata extraction. AMeGA addresses the issue that extraction based on visual or linguistic characteristics results in local solutions that cannot be adapted more generally. They also address the issue that the document creators (individuals) usually have no professional training in metadata creation [61]. AMeGA says that: "Experimental automatic metadata generation research projects have had little focus on using the documents generated by content creation software as a data source." This is not particularly accurate since there were several projects using harvested metadata from HTML documents. However, it is still correct for other document formats, such as PDF and MS Office document formats.

Based on this background information, professional metadata labellers were interviewed regarding their trust in automatically generated metadata based on the Dublin Core schema. As would be expected, AMeGA discovered data indicating that trust in AMG depended upon the specific schema element. AMeGA concluded that this user group would like to use automatically generated metadata, although 96% wanted the ability to make corrections after the AMG process. The AMeGA project recommended evaluation of metadata sources by using statistical data, and the use of an external data source for creator information.

Syn et al. extended AMeGA's efforts by using the Metadata Generation System (MGS) to analyse metadata gained from harvesting Meta-tags from HTML documents, using different linguistic-based rules for subject and keyword extraction, plus descriptions of sub-components that made up the webpage [128]. They present the user with automatically generated entities, from which the user chooses in generating the metadata record for the webpage. They concluded that such a service can generate metadata of quality equal to manually generated metadata, and that such methods have special value when the user is not a professional metadata labeller.

Jorum project

The Jorum project developed an online repository service for teaching and support staff [14, 84]. Contributions could be made in the form of imported IMS Manifest content packages or as stand-alone documents. The IMS Manifest content packages contain professionally created metadata descriptions based on the IEEE LOM schema. These entities are harvested by their system.

Stand-alone documents go through multiple AMG efforts, with a title generated based on the file name. The researchers discussed the filtering of document format extensions such as ".doc" as an option for future research. The system generates entities for the "Identifier"; the automatically set date based on the system time; default entities for elements including "Language" ("EN"), "Role" ("Creator"), "Metadata schema"

("LOM 6.2," "IMS 1.2.1," "JORUM"); vCards for publisher information; the "Format" based on the document format extension (which does not include the document format version); the "Size" (collected from the publishing system); and "Rights" based on default entities set by the institution associated with the user specified the user profile. Metadata usage (changes) is not registered. The user must manually specify in the meta-metadata when he or she has made metadata corrections.

Jenkins et al.

Jenkins et al presented a project where they automatically generated 10 of the 15 metadata elements from the Dublin Core metadata schema [82]. This was done for HTML documents. Their main efforts were on generating a range of metadata elements and not specific elements. However, there were more extensive efforts performed to generate entities for the "Subject" and "Description" elements.

The "Title" elements were harvested from the HTML Meta-tag. They used the HTTP protocol to harvest data regarding the targeting document, including "Date," "Format" (content) (MIME type, not file format version), and "Format" (extent) (the file size). They used rule-based algorithms based on natural language to specify keyword stored as the "Subject" element. They counted unique words and removed common stop words. Their algorithm gave words from the header extra credit and used words with the highest 10% of scores as the "Subject." The first 25 words found in the body of the document were used as the document description. Hyperlinks were extracted to generate "Relation" entities. The number of document pages was used as a proxy for download time [82].

Additionally, the system provided entities for the "Identifier" element based on the URL. "Rights" and "Publisher" information were collected from a configuration document on the system server. The entity for the "Creator" element was collected from a document containing those entities that was located on the user's home directory to the system. Jenkins et al. presented this as an easy and not particularly accurate solution for generating "Creator" elements [82]. Jenkins et al. did not describe how this model functioned in terms of a correctness rate [82].

In an earlier project, [83] showed automatic classification of HTML documents in accordance with the Dewey Decimal Classification system (DDC) [115]. They used a combination of harvesting and extraction to generate metadata. The "Title" and "Description" elements were collected from meta-tags. If no "Description" tag was present, then the first 25 words were used. Keywords were generated based on comparisons between the words of the document and terms found in the representatives for various DDC classes. The HTML documents were also classified based on their DDC keywords. The "Date" (modified) element was collected from the HTTP header. The classification date was set based on the system time. Word counts were conducted on all written content in order to describe the document extent, detail, and download time [83].

Hlava

Hlava describes two approaches for AMG of keywords [70]:

- **(Statistical) Machine Learning:** Use of the principle that if two words occur frequently together, then those words are related conceptually. To perform such a task, various algorithms have been used. Hlava continues by citing how large projects have not been able to generate metadata of sufficient quality in order to serve their purpose even with thousands of documents in which the used algorithms can perform "training" upon [70]. With a typical error rate of 40 to 60 present, the practical usefulness of the generated metadata is limited. An extensive amount of research is currently conducted based on Machine Learning in order to generate keyword, subject and context metadata, e.g. Heidorn et al. [67] and Kim et al [86].
- **Rule-Based:** This approach uses a pre-set list of terms; typically a thesaurus, taxonomy or authority file. Hlava states accuracy of about 60 present for "simple" algorithms that compare an input against their reference, and 85 to 95 present for "complex" rules where additional human included logics have been included in the algorithms [70].

This thesis have in many sense used efforts described above as "complex", Rule-Based. This since harvested document contents have been compared based on pre-defined rules with additional logics, such as content filtering, in order to achieve higher quality metadata. However, as article P5 presents, higher quality metadata results can be achieved using the "simple" Rule-Based approach, which is not limited by human rule generation efforts, if accurate reference terms are present [P5]. By using University course specific descriptions as reference terms, published documents from courses could be given high quality metadata with only a need for simple, domain specific filtering.

Search engines

Commercial search engines use AMG algorithms to generate metadata for local resources and for content on the internet. The algorithms which actors such as Google, Yahoo and Microsoft use in their search engines are trade secrets. As such, it is not possible to point to specific issues which have been addressed over the last few years. However, it is possible to compare the results of these algorithms when performing identical queries. This thesis has been reviewing search engine performance over the last years without any major improvements in terms of the quality of the data presented in the search results.

Google

Google harvests the "Title" element from HTML and MS Office documents [57]. No filtering is performed to exclude false harvested metadata. Google has been using extraction to generate the "Title" element from PDF document. The company's extraction algorithm is a trade secret. This researcher has observed that the Google algorithm focuses on extracting only the words with the largest letters on the first page for the "Title." The "Title" element is the only element presented aside from the

document content in query results. More recently, Google has moved away from extracting titles from PDF documents and is now harvesting the "Title" element from PDF documents as well [58]. No filtering or quality enhancement seems to be performed on the metadata.

Google Desktop uses other algorithms: Here the “Title” element is harvested from PDF documents [59]. If no “Title” element is embedded, then the file name and possibly file path (depending on the length of the file name) is presented as the title. The “Title” element is harvested from e-mail messages made available through the MS Office-based application Outlook, although, titles from MS Office document formats are not harvested. Instead these documents are presented with their file name and possibly their file path.

Yahoo

Yahoo harvests the “Title” element from HTML documents [135]. Yahoo does not present a title for MS Office documents. Instead the web address is presented at the location where the “Title” is normally presented in the graphical user interface (GUI). Yahoo extracts the title from PDF documents. Their extraction algorithm is a trade secret. This researcher has observed that their algorithm focuses on extracting only the words with the largest letters on the first page for the “Title.”

Scirus

The scientific search engine Scirus uses HTML tags to harvest the “Title” element from HTML documents [121]. This service does not analyse Word, PowerPoint and Excel document formats. PDF document titles are harvested. If no “Title” element is embedded, then all text on the first page is presented as the title. No filtering is performed to exclude false harvested metadata. The “Title” element is the only element presented aside from the document content in query results.

Rule-based approach

In this third of four subchapters, we look at various research systems and projects which have explored usage of rule-based AMG efforts. These are typically smaller scale research efforts exploring generation of specific metadata elements using rule-based AMG efforts. First up are Flynn et al. and their usage of visual characteristics to classify documents.

Flynn et al.

Flynn et al. used rule-based visual characteristics to classify documents based on their visual appearance compared to category specific templates [49]. By first classifying each document into a specific document type, they demonstrated the use of rules based on visual characteristics that were fine-tuned to the specific document template. In this manner, more rules and more precise rules could be executed without having these rules generate false results for documents with another type of visual appearance. By doing this, this project managed to increase the correctness rate of the “Title” and “Author” elements.

LOMGen

The Learning Object Metadata Generator (LOMGen) project presented extraction of metadata from highly structured HTML documents in order to generate CanLOM metadata [98, 123]. They proposed using meta-tags for generating the elements "Title," "Description" and "Keywords," and use of a rule-based natural language algorithm to generate classification data. Results from their use of this approach have not been published.

MAGIC

The MAGIC (Metadata Automated Generation for Instructional Content) system generated metadata for SCORM objects [92]. These objects can consist of content in the form of audio, video and text data sources. The researchers used a range of different algorithms to generate textual content from these data sources. Graphical text documents, such as PDF and Word documents, are converted to plain text. MAGIC uses rule-based algorithms based on natural language for generating "Title," "Keywords" and "Summary" entities. Their exact method is not published. This project shows the potential of using AMG to generate metadata for audio- and video-based documents in addition to plain text.

Kawtrakul et al.

Kawtrakul et al. demonstrated use of the rule-based approach based on visual characteristics [85]. They used documents with a well-defined structure as a data source. Their rules were then adapted to the local dataset. They showed extraction of their documents' headers, consisting of a range of pre-made elements with a pre-defined visual structure:

```
<author-name> <year> : <thesis-title> . <degree-name>,
<major-name>, <department-name>. <advisor-name>,
<advisor-degree>. <page-number> pages.
```

Kawtrakul et al. stated that such a solution needs to be adapted to the local visual document structure to perform optimally [85].

Liu et al.

Liu et al. also applied the rule-based approach based on visual characteristics [97]. They used rules to locate tables within documents and for extraction of the table content.

Boguraev et al.

Boguraev et al. used rule-based algorithms based on natural language [16]. They employed lexical repetition to generate linguistically aware summaries of articles from The New York Times. They used rules based on visual characteristics to identify sections in which content was presented.

Giuffrida et al.

Giuffrida et al. presented metadata extraction efforts based on scientific papers as a data source [56]. These were collected from conference and journal papers that were

published in the PostScript document format. They used rule-based algorithms based on spatial visual characteristics to identify the document content, as in:

> *"The title is located on the upper portion of the first page and it uses the largest font on the first page; Authors are listed immediately under the title in a certain order; Affiliations follow the authors' list;…"*
>
> Giuffrida et al. [56]

These rules were constructed based on the project authors' knowledge of the dataset formatting. This project managed to receive correctness rates of 92% for "Title," 87% for "Author," 75% for "Affiliation," 71% for multiple "Affiliations," 76% for table of contents. To achieve this, they used 9 title rules and 12 author rules. However, the citation above is the only actual rule that was presented.

Machine learning approach

Finally we look at AMG efforts utilizing the machine learning approach. These efforts commonly strive to generate entities to one element or a small selection of elements. Compared to the rule-based approach projects, these machine learning approaches include a higher grade of complexity or execution logics.

First up is Liddy et al. which stands out from the crowd of machine learning approach efforts by attempting to generate a range of entities for their metadata elements.

Liddy et al.

Liddy et al presented the use of machine learning to generate metadata following the Gateway to Education Materials (GEM) metadata standard [53, 94]. This project showed that a range of metadata elements can be automatically generated. They used natural language-based rules to generate entities for all the elements of the GEM schema. They used existing lesson plans from their collection of pre-registered GEM documents as a data source. Sets of metadata descriptions were compared to manually created metadata records. The actual results from the automatically generated metadata were not presented, although the researchers stated that reviewers regarded these records to be comparable to the manually created records, scoring roughly 10% lower on a locally used scale for measuring satisfaction and expectations. However, the adaptations made to accommodate their dataset made their solution unsuitable for general learning object metadata registration.

Seymore et al.

Seymore et al. extracted metadata based on the heading of research papers as a data source [122]. They explored using the machine learning approach by making use of hidden Markov models (HMM) for the information extraction tasks. Han et al. used the same dataset for analysis of using an alternative model called the Support Vector Machine (SVM) [66]. Takasu used Optical Character Recognition (OCR) to extract content from bibliographies of books and academic articles [130]. Takasu used a Dual and Variable Hidden Markov Model (DVHMM) to extract metadata from the data source.

All these models used the visual characteristics of their data source to "teach" their system how to recognize different elements. The systems were used to generate metadata elements for title, author information, abstract and keywords. The results of all these projects all showed the potential found in such machine learning technologies. But the studies also demonstrated the major weakness with machine learning, which is the reliance on a pre-known, well-structured data source. These models were adapted to work in a specific context with close to standardized, very structured and strictly formatted scientific papers. Placed in a more general context where there is less common visual document appearance structure, these models will not provide quality metadata. There is a need for more generally valid tools for AMG.

Xue et al.

Xue et al. argued that the "Title" meta-tag in HTML documents is seldom representative as a title for the document [134]. They used of machine learning to generate "Title" elements. They presented the use of the Direct Object Model (DOM) tree for generating a formatted dataset along with two different models for metadata generation: The Support Vector Machines (SVM) and Conditional Random Fields (CRF) models. They showed promising results, although they also expressed how vulnerable machine learning models are to changes in the dataset. Results from using a standardized test-dataset showed a correctness rate of between 11% and 64% depending on the sub-section of the dataset. Xue et al. expressed a need for combining data sources to obtain better correctness rates [134].

Hu et al. presented usage of machine learning based on visual characteristics [71]. This project used a range of different algorithm models to generate metadata. This project reported a correctness rate of 83.7% for Word and 89.5% for PowerPoint[4]. They presented the document format features as the key to successful title extraction. They investigated the use of linguistics as a data source instead of visual characteristics. They concluded: "It does not work well. It seems that the format features play important roles and the linguistics features are supplements." An error analysis showed the reasons why errors occurred:

a) One-quarter were caused by documents without a "true" title.

b) One-third was caused by documents with layouts that were difficult to understand.

c) The remaining (about 42%) was caused by confusion regarding titles and sub-titles.

The issues stated in (b) and (c) can be addressed by analysing the document format.

In earlier work, this research group used the Support Vector Machines model [91]. They constructed a system for categorizing content from an Intranet to enable queries that

[4] Word: Precision = 81,0%, recall = 83,7%. PowerPoint: Precision = 87,5%, recall = 89,5%.

distinguished between definitions, persons, experts and homepages. This service relied on having "Title" and "Author" metadata to make these distinctions. Li et al. examined Word and PowerPoint documents collected from Microsoft's own Intranet systems [91]. These contained a greater number of visual differences than other datasets consisting of scientific papers. But these documents should still be regarded as being structured in a similar manner as compared to the general situation. They reported the correctness rate of the embedded "Title" element to be 26.5%, while the rate for the "Author" element was 12.6%. They used a machine learning and Support Vector Machine model for visual content analysis. Regarding their "Title" element, they achieved a correctness rate of 89.9% for Word and 95.1% for PowerPoint documents[5]. The results of their "Author" algorithm were not published.

Li et al. [93] undertook a project that was very similar to [82]. Li et al. also generated 10 metadata elements based on the Dublin Core schema [93]. Nine of these elements were generated in the same way. The exception was the "Subject" element that was generated by using another natural language rule-based algorithm. Their "Neural Network" algorithm used stopping, stemming and weighting of the document content. Stopping removed high frequency words with low content discriminating power, such as "to," "a," "and" and "it." Stemming was used to reduce the document content to only "root words." This meant correcting the words "compares," "compared" and "comparing" to the root word "compare." The weighting was performed using two different models. Here the "EFT-IDF" model counted the number of times a word was repeated in a document relative to the number of words in the document. This list of words was compared to a total list of words in the specific dataset. The words that occurred least frequently were assumed to be of greater importance in order to distinguish between documents. The alternative "PCA" model was used to "increase feature variation and decrease feature space dimensionality."

2.3.6 Summary

To sum up, each of the AMG algorithm approaches described above has its own strengths and weaknesses:

- **Harvesting:** Uses data that are easy to access and collects entities from embedded metadata stored as part of the document code. This approach's main weaknesses are: (1) the limited amount of elements in practical use; (2) uncertainty about whether the elements selected contain entities that reflect the document; and (3) because the number of people who are aware of embedded metadata is so limited, few people work on generating and correcting this metadata.
- **Extraction using rules based on visual characteristics:** This approach can be used to identify and collect a large number of elements. Its intent is to collect content specified by the user. This approach's main weaknesses are: (1) its requirements regarding knowledge of the documents used: (2) the requirement

[5] Word: Precision = 87,5%, recall = 89,9%. PowerPoint: Precision = 90,7%, recall = 95,1%.

for standardized formatted documents; (3) the possibility that it will require a labyrinth of rules that need to work together; and (4) issues regarding multiple candidate elements.

- **Extraction using rules based on natural language:** This approach has the potential of generating semantic metadata, such as classification, subject, keywords and description, which even humans can find difficult to generate. However, its main weaknesses are: (1) it requires extensive knowledge of the document contexts, limiting it to specific subjects and specific languages; (2) it is limited to specific elements, requiring it to be used along with other metadata efforts for practical usage; and (3) it does not scale to a general purpose context or a multi-linguistic environment.
- **Extraction using rules based on the document code:** This approach can be used to collect all user-specified content from template sections regardless of visual document presentation or the language of the intellectual content. It enables blank AMG results if no section content is collectable, avoiding the generation of multiple candidate entities. It can be used to collect document descriptions that are part of the document code, such as references, language tags, illustrations and tables, and can provide extensive descriptions of the document, which in turn can be used to increase the correctness of other AMG algorithms. This is accomplished by providing a data source based on facts rather than software based on judgment and by providing direct access to the main document content. Using this approach has the following drawbacks: (1) requires extensive knowledge of the specific document format and templates used; (2) requires extensive knowledge of how applications use the document format and template; (3) is vulnerable to misuse if a new application or application version uses the document format in an unanticipated manner; (4) access to document content is difficult for humans to understand, because of the binary document formats; and (5) there are very few scientific efforts at present.
- **Document context analysis:** This approach can be used to generate default elements that are correct for most of the documents in a collection, and to generate element types that are not commonly harvested or extractable, such as educational elements. The use of this approach requires that: (1) the default entities be actively corrected to sub-collections in order to provide value; and (2) the user recognize that the approach is not adapted to describe the individual document, rather collections of documents. Hence less accurate metadata may be among the results.
- **Document usage:** This approach can be used to gain knowledge of *actual* usage instead of *intended* usage and to gain knowledge of usage patterns that do not have other potential AMG data sources, such as typical learning time for papers. The use of this approach requires: (1) registration and logging of user actions (surveillance); and (2) that the LMS be used for "all" document related usages. Once a document is used outside of the LMS, then the LMS is not able to register the document usage.

- **Composite document structure:** This approach allows new documents to obtain metadata from existing documents that are closely related. However, it requires that the first version be generated manually or by using other AMG methods. This limits the extent of systems where such algorithms would have practical effects, because few documents within the observed LMS have been re-published. Instead they are replaced by new documents.

2.4 Learning Objects

2.4.1 Introduction

We now have a framework for specifying quality and tools to perform AMG. And we need a document collection in which we can validate various AMG effort results. This thesis therefore went on a hunt to locate the best document collection in order to perform its analysis. There were a number of candidate document collections including:

- **Library records:** Very strictly formatted documents which metadata should follow a strict and limited metadata schema. This gives the AMG algorithms less of a challenge, as it is known what the AMG algorithms should look for and where the desired content is located; the document diversity is missing. However, there are a number of document collections available to perform research upon.

- **Medical records:** Very strictly formatted documents which metadata should follow a strict but larger metadata schema. This can give the AMG algorithms more challenges to place located entities, though the AMG algorithms should to a large extent still know what to look for and where the desired content is located. Here too is the document diversity missing, and there are a number of document collections available to perform research upon.

- **Academic publications:** Very similar to medical records in terms of metadata schema complexity and visual appearance of the publications. As with library and medical records, the document diversity is missing, and there are a number of document collections available to perform research upon.

- **Documents on a company intranet:** The guidelines vary from company or organization, though commonly documents following a guideline published without or with limited metadata. Here we find the desired document diversity, but missing the metadata schema to populate with entities. Document collections are also not freely available.

- **Documents on an open educational network:** By choosing an educational network where documents for a number of subjects are shared, and where the users actively share "whatever they want", we get the desired document diversity. And for educational documents there are a number of highly detailed educational metadata schemas which could be populated with entities and allow us to explore the various types of AMG algorithms. However, gaining access to

such documents can be a challenge, as pre-made document collections are not present.

Of the various candidate document collection types, educational documents from an educational intranet were chosen in order to gain access to highly diverse documents in terms of visual characteristics, technical formatting, subjects and even language of the intellectual content. And there are usages of documents and their metadata descriptions that can be combined in order to enable sharing of document with their metadata descriptions – A task very seldom seen, as manual generation of the educational metadata takes an extensive amount of time. On the other hand gaining access to such documents can be challenging since the documents need to be located, retrieved and made into a document collection. In addition, the extreme document diversity is an extraordinary challenging for AMG efforts, which could be one of the reasons why AMG research efforts seem to avoid diverse document collections.

This thesis sticks to documents from an educational network in order to explore the possibilities and limitations of various AMG algorithms. Documents that in this educational context is commonly referred to as Learning Objects (LOs).

With the presence of descriptive metadata there is a potential of efficient document sharing and retrieval without having to inspect each and every document each and every time a query is performed. This potential is however only available if:

1) The desired metadata is registered
2) That these descriptions are correct
3) That the metadata descriptions are understood correctly by the querying application or person.
4) That the document is available!

The first issue in the list above is addressed in Chapter 2.3 with its presentation of techniques to automatically generate descriptive metadata using various data sources to create a range of different metadata.

The second issue is addressed in Chapter 2.2 with measures to ensure that the quality of the metadata is sufficient.

The third issue reflects on what metadata descriptions that are desired generated and how the generated metadata should be stored. Metadata *schemas* are used for such tasks by giving a description of what and how such metadata should be registered and the intentions for doing so. There are a number of different metadata schemas created to standardize metadata within one or more fields. Chapter 2.4.3 reflects upon this by presenting various metadata schemas, from one of the most general standards (The Dublin Core) to subject specific standards including ADN and SCORM.

The value of metadata is reduced dramatically if the objects which they describe are not available. Chapter 2.4.5 reflects upon a real-world, large scale system where documents

are published and where tens of thousands of users try to locate *the* right documents for them. This chapter describes the NTNU Intranet called It's learning, which is used for sharing educationally related documents at the University, without any restrictions to document contents or formatting (besides an upper size limit). Hence this is a prime example of a system with an extensive range of document types and subjects made available, and where the documents should be described using the same metadata schema in order to enable discovery of documents using search methodology. By using this system as a test system this thesis gains access to a large number of non-homogeneous documents made available for sharing. This thesis will work with such documents and AMG methodology in order to generate standardized metadata in accordance with metadata schema with metadata entities in accordance with quality goals.

This thesis is not limited to "documents". A more precise definition of the objects or resources that this thesis will be working with is given in Chapter 2.4.2.

Chapter 2.4.4 presents how standardized metadata and published educational resources can be combined in order to enable sharing of educational resources with rich technical and educational metadata descriptions.

2.4.2 Defining "Learning Object"

The field of digital educational documents is relatively new. This is reflected in the many different expressions used to identify these types of documents and in the many definitions that have been presented. The literature is full of suggestions for names for these documents, such as: "Knowledge objects" [102], "Components of Instruction" [110], "Pedagogical documents" [10], "Educational software components" [45], "Online learning materials" [101], "Documents" [6], "Instructional components" [119] and "Learning object" [72]. From this list of candidate phrase names, the expression "Learning Object" has become a standard phrase due to the adoption of the IEEE LOM metadata schema, presented in Chapter 2.4.3. The IEEE Learning Technology Standards Committee (LTSC) defines a Learning Object as:

> *Any entity, digital or non-digital, that may be used for learning, education, or training*
>
> IEEE LTSC [73]

This definition covers content such as curriculum lists, personal lists, notes, books, printed and digital articles, presentations and multimedia elements, to name just a few. By using such a broad definition, learning objects in all educational subjects can be described with this model. This definition does not differentiate between candidate Learning Object types, in that there is nothing in the definition that describes what a LO can be and how extensive it can be. In this sense, all the content listed below should be regarded as individual LOs:

- The whole LMS with all its content.
- Each course section with all its content.
- The document created within the course section.
- The stand-alone documents uploaded as part of a system specific document type.

We need to examine the idea behind the concept of a LO in order to create boundaries for where one LO stops and where another begins. There is an agreement regarding LOs that they should be components that can be put together to create the educational material needed for a teacher to perform specific educational tasks. In this sense, the LMS should still to be regarded as a LO. However, the LOs need to be set in an educational context to identify the LOs most suitable for the task at hand. As there is the possibility to use far smaller components as an educational document, the primary focus of an LO should be on smaller components of data. By going to the specific educational section, there is a still an extensive need to segment the data source in order not to mix up content that does not logically belong within a single object. Wiley developed taxonomy for defining five different LO types: Single-type LO, Combined-intact LO, Combined-modifiable LO, Generative-presentation LO and Generative-instructional LO [133].

Table 4: Taxonomy of Learning Object Types, based on [133]

LO characteristics	Single-type LO	Combined-intact LO	Combined-modifiable LO	Generative-presentation LO	Generative-instructional LO
Number of elements combined	One	Few	Many	Many – Few	Few – Many
Type of object contained	Single	Single, combined-intact	All	Single, combined-intact	Single, combined-intact, generative-pres
Reusability of component objects	(not applicable)	Low	High	High	High
Common function	Exhibit, display	Pre-designed instruction or practice	Pre-designed instruction and / or practice	Exhibit, display	Computer-generated instruction and / or practice
Extra-object dependence	No	No	Yes	Yes / No	Yes
Type of logic contained in object	(not applicable)	None, or answer sheet based item scoring	None, or domain-specific instructional and assessment strategies	Domain-specific presentation strategies	Domain-independent presentation, instructional and assessment strategies
Potential for inter-contextual reuse	High	Medium	Low	Low	High
Potential for intra-contextual reuse	Low	Low	Medium	High	High

The individual documents created within the case LMS would therefore be regarded as "Single-type Leaning Objects," since these are the smallest LO components within the system. Stand-alone documents should be permitted as LOs, according to the LO definition. However, the case LMS does not allow stand-alone documents to be published outside of an existing LO. The LOs that contain attached document(s) could therefore be regarded as either "Combined-intact LOs" or "Combined-modifiable LOs." In order to avoid inconsistencies regarding terms, this research uses the phrase "document" to refer to LOs and other resources that are discussed in this thesis. The phrase "LO" is only used in this and the following chapter.

2.4.3 Learning Object Metadata schemas

Metadata schemas are systematic descriptions of which metadata elements which can be presented and how these elements can be presented.

A wide range of metadata standards is in use today. Of these, the Dublin Core (DC) standard is the most widely used and in many communities serves as a general de facto standard for describing the global properties of objects, including the title, creator and subject [28]. Other metadata standards have been developed with a finer grain than the DC's 15 metadata element structure, making them able to give more precise and specific metadata descriptions. In terms of learning object description, the IEEE LTSC [75], IMS [79], ARIADNE [10] and NSF [113] are major players who have made important efforts to provide a finer-grained tool to describe learning objects. These efforts include information about how to use a learning object, where it has been used and what kind of learning material can now be stored. What we see today is a trend of merging standards in order to share a fine-grained and structured metadata schema, build exchange capabilities between systems without losing fine-grained functionality, and to increase their functionality. This chapter gives an overview of different educationally related metadata schemas, and the contexts in which they are employed.

Dublin Core

A wide range of metadata standards is in use today. Of these, the Dublin Core (DC) metadata schema standard is probably the most widely used and in many communities serves as a general de facto standard for describing the global properties of objects, including the title, creator and subject [28]. The DC metadata schema [28] is designed to be a simple metadata standard [27]. It has been designed to let document authors and publishers to create metadata descriptions without requiring pre-training. It offers a metadata schema for describing 15 basic and commonly used metadata elements. These are all recommended elements that can be used if the author or publisher chooses to.

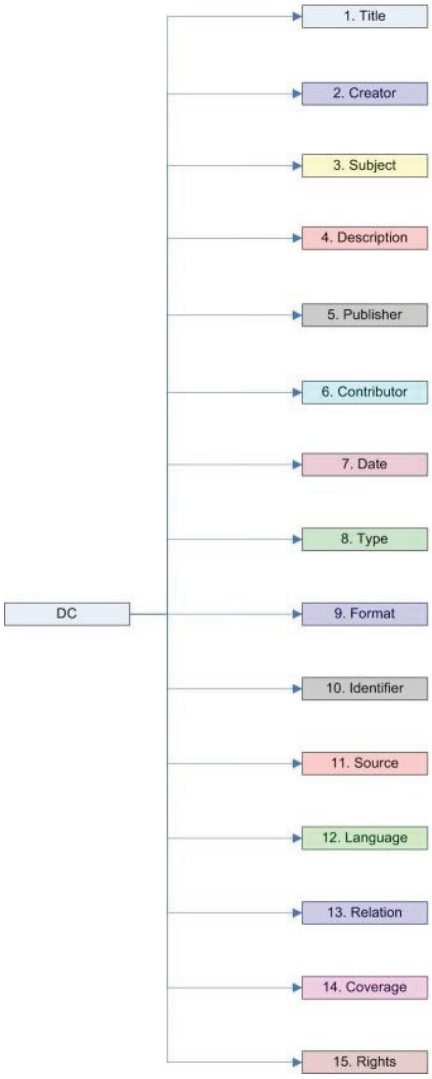


Figure 19: The Dublin Core schema

The DC schema offers the use of "Qualifiers" that allow for more detailed metadata descriptions [24]. Qualifiers are used for two purposes: (1) To specify encoding schema(s), e.g. standards for presenting dates (such as "DD.MM.YYYY"), and (2) to specify element refinement(s) (e.g. the "Date" element can, with qualifiers, be extended to describe the "Created," "Valid," "Available," "Issued" and "Modified" dates).

DC has been widely adapted by actors in the computer industry. DC metadata have been included as part of the metadata schemas of commonly used document formats, such as MS Word and Adobe PDF. It is also compatible with a number of different metadata schemas. However, since the schema is quite basic, a number of projects have been initiated to expand the schema to local and subject-specific needs.

Dublin Core Extension

One such effort is in describing Learning Objects. The original DC metadata schema lacked basic elements for describing document use in an educational context. An extension of the original schema has therefore been developed, called the DC-ed (DC

educational) [25]. This extended schema includes metadata elements for: “Audience,” “Instructional Method,” “Conforms To,” “Education Level” and “Mediator.”

In addition, this schema extension includes requirements regarding obligatory elements. The DC-ed is a project under development. Efforts are underway to extend the compatibility between the DC-ed and an even more extensive metadata schema, the IEEE LOM [26].

Table 5: Timeline of educational initiatives and standards

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
DC initiative started			DC v1	DC v1.1						
				IMS v1		v1.1				
							IEEE LOM D1			D8
		ADN started					ADN			v.0.6.50
ARIADNE started										
							NSDL			
		GEM v1							V2	
									DC-ed	DC-ed
	European Schoolnet started									v3

The metadata schema Gateway to Education Materials (GEM) also uses DC elements plus 8 educational elements [53]: “Audience,” “Instructional Method,” “Cataloguing,” “Duration,” “Essential Documents,” “Provenance,” “Rights Holder” and “Standards.” There are common elements among the extensions made by DC-ed and GEM. However, GEM has a bit of a different focus; the DC-ed describes the document itself, whereas GEM describes more of the educational context of the document.

The National Science Digital Library (NSDL) has also generated an educational metadata schema based on DC [112]. They have extended their schema to include the element “Audience,” which is an element that has since been included in DC-ed. In addition, they have included the original IMS meta-data schema standard elements of “Interactivity Type,” “Interactivity Level” and “Typical Learning Time,” which are now also present in the IEEE LOM schema.

The European Schoolnet developed their EUN metadata model based on the DC with additional elements for rights, approval, release, user level and version [37, 95]. The European Schoolnet’s model has over time evolved to include more elements and mandatory elements. In its latest version, their metadata schema, now called the Learning Document Exchange Metadata application profile (LRE AP), is presented as based on the IEEE LOM [48]. Several national initiatives for learning object creation and distribution are based on using the European Schoolnet standards, such as the Norwegian Skolenettet [124] and the Swedish Skolnet [47].

Even though "Creator" is the element name commonly used in metadata schemas for identifying the user that has created the document, it is not the name used in PDF, Word and PowerPoint document schemas. Here the element name "Author" is used instead. In PDF documents, the "Creator" element is reserved for the creator application name rather than person name. PDF documents can contain multiple sub-schemas, such as sections containing the elements "DC. Creator" and "XAP. Author." These elements can be used as alternative data sources within PDF documents. Word and PowerPoint documents can contain a "Last Author" element with metadata, which describes the last person to edit the document. Using harvested, existing metadata requires knowledge of the document format's schema and the destination schema.

IEEE Learning Object Metadata

On the other side of the scale we find the complex educational metadata standard IEEE Learning Object Metadata (LOM) and its "cousin" the ADN. These are extensive and complicated metadata schema. Both describe so-called "Learning objects" or as the mother organization IEEE LTSC states;

> *Any entity, digital or non-digital, that may be used for learning, education, or training*
>
> IEEE LTSC [73]

In the context of this thesis, LOs are documents published on the NTNU Intranet, called "It's learning".

IEEE Learning Object Metadata Extension

The IEEE LOM is an extensive and complicated metadata schema. It has 9 different metadata element category branches, where each category includes between 3 and 11 elements. There are 45 basic elements, although a number of these elements have sub-elements that make them suitable for multiple usage areas.

The IEEE LOM has been developed by the IEEE LTSC [75], IMS [79] and ARIADNE [10]. These actors have been working to build an extensive metadata schema for detailed educational descriptions. The standard is based on the IMS metadata standard from 1998, which was accepted as an IEEE specification, known as IEEE LOM in 1999 [78]. Since then the standard has gone through several versions. The main structure has remained unchanged, though selected elements and specific value spaces have been added. The use of elements has also changed, allowing more element descriptions and eliminating obligatory elements.

The IEEE LOM schema is used as the metadata schema for the Sharable Content Object Reference Model (SCORM) [2] specification of the Advanced Distributed Learning (ADL) Initiative [3].

The IEEE LOM has the support and potential to be the standard for learning object metadata exchange in the years to come, making it a central point when studying learning object metadata standards. A variety of local, national versions of the IEEE LOM has been or is under development, including the UK LOM [22], NORLOM [46],

SWELOM [51] and CanLom [19]. Friesen and Neven et al. present a range of different LMSs or Learning Object based Repositories (LOR) in a survey of LOM-based repositories [52, 108].

The IEEE LOM allows the creation of meta-metadata, which are metadata that describe both the metadata contributor (generator) and specific elements and entities. This enables the inclusion of multiple identical elements with different entities without logically corrupting the metadata records. Such differences can occur when users have different opinions regarding the document. Meta-metadata enable the use of the metadata records as feedback channels from the document users. Such metadata are called “non-authoritative” metadata [118]. Metadata published by the document creator or publisher are called “authoritative” metadata. The IEEE LOM schema does not provide elements that distinguish between actual and intended usage, the number of pages, and the number of characters, and does not provide any assessment of the document quality in the form of the publisher’s role. This schema does not provide elements for distinguishing between publications by students, lecturers or other administrators. The IEEE LOM lets the user create custom elements in its “Classification” section. Though, this can result in compatibility issues with other systems that use the IEEE LOM schema. The IEEE LOM has proven to be amenable to change if new ideas are presented that could provide additional functionality that is currently not supported. One example of this is the GESTALT project [54], which needed increased functionality for its users [50]. They then added new technical metadata, and a much more detailed metadata structure for describing personal contact information, including e-mail, post address, fax and telephone. These changes have since been incorporated into the IEEE LOM schema.

Possibly the most major drawback of using this schema is that it requires extensive user pre-training and that it is labour intensive to generate metadata records [92]. IEEE LOM metadata records are known to take more than one hour to manually label [52].

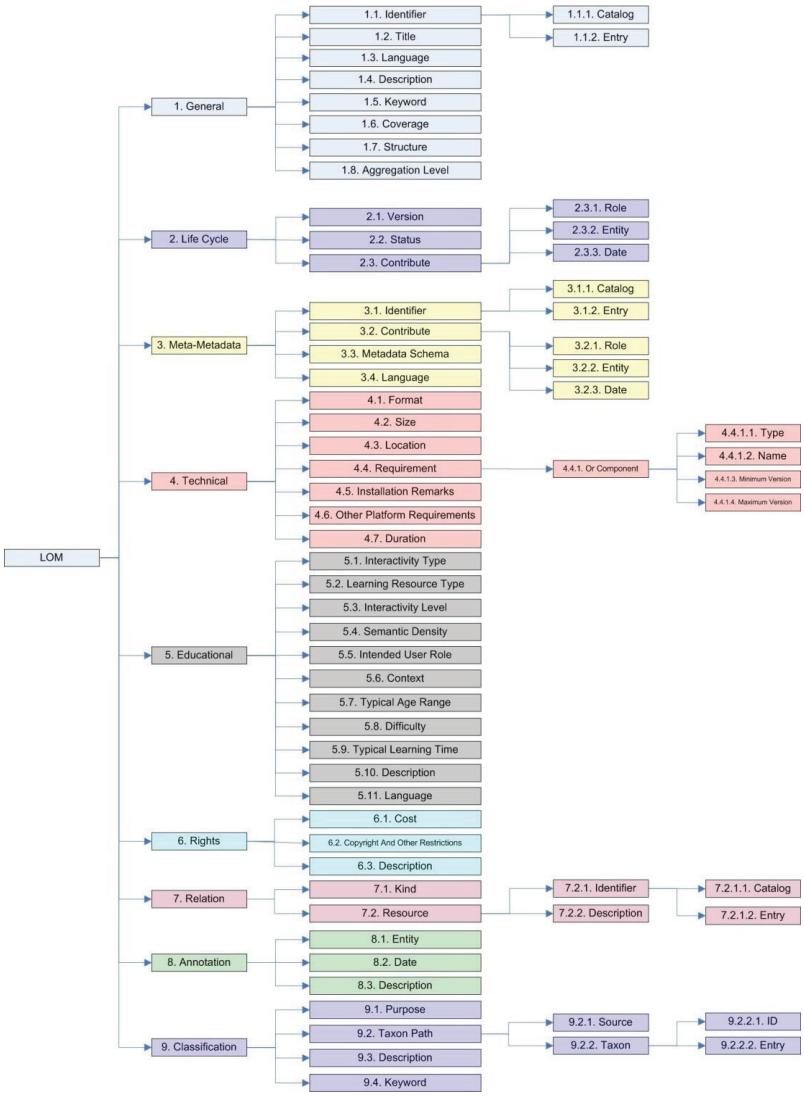


Figure 20: IEEE LOM (Draft 8)

ADN

The other major player in the development of digital repositories for educational documents is the American NSF with their Digital Library Initiative (phase 2) [34]. NSF decided to streamline their metadata structures by merging the existing metadata standards previously used by their projects. ADEPT [1], DLESE [33] and NASA Earth Science Enterprise [107] therefore combined efforts in 2001 on a project called ADN [32]. This metadata framework was created to describe educational documents in the American educational system and to fully comply with the requirements of all involved

agencies [32]. ADN is based on the same IMS standard as IEEE LOM. However, the development of these two schemas has taken different approaches in regards to extendibility; The ADN schema is highly adapted to its specific use and uses an extensive network of sub-sections in an object-oriented manner, whereas the IEEE LOM allows local adaptation for increased usability with less defined custom extensions.

The ADN presents itself as a stricter schema that is highly adapted for the exclusive use of professionals for metadata creation on a national level. The extensive use of elements also makes use of the schema for those querying documents more challenging since it is difficult to specify and interpret this level of detail without professional knowledge of the subject. The ADN schema cannot be regarded as usable in a general educational context or outside of the American educational system. As a consequence, American projects commonly employ the IEEE LOM schema instead of the ADN, including the efforts of the DC Educational Community [26]. This thesis has hence focused on IEEE LOM rather than ADN.

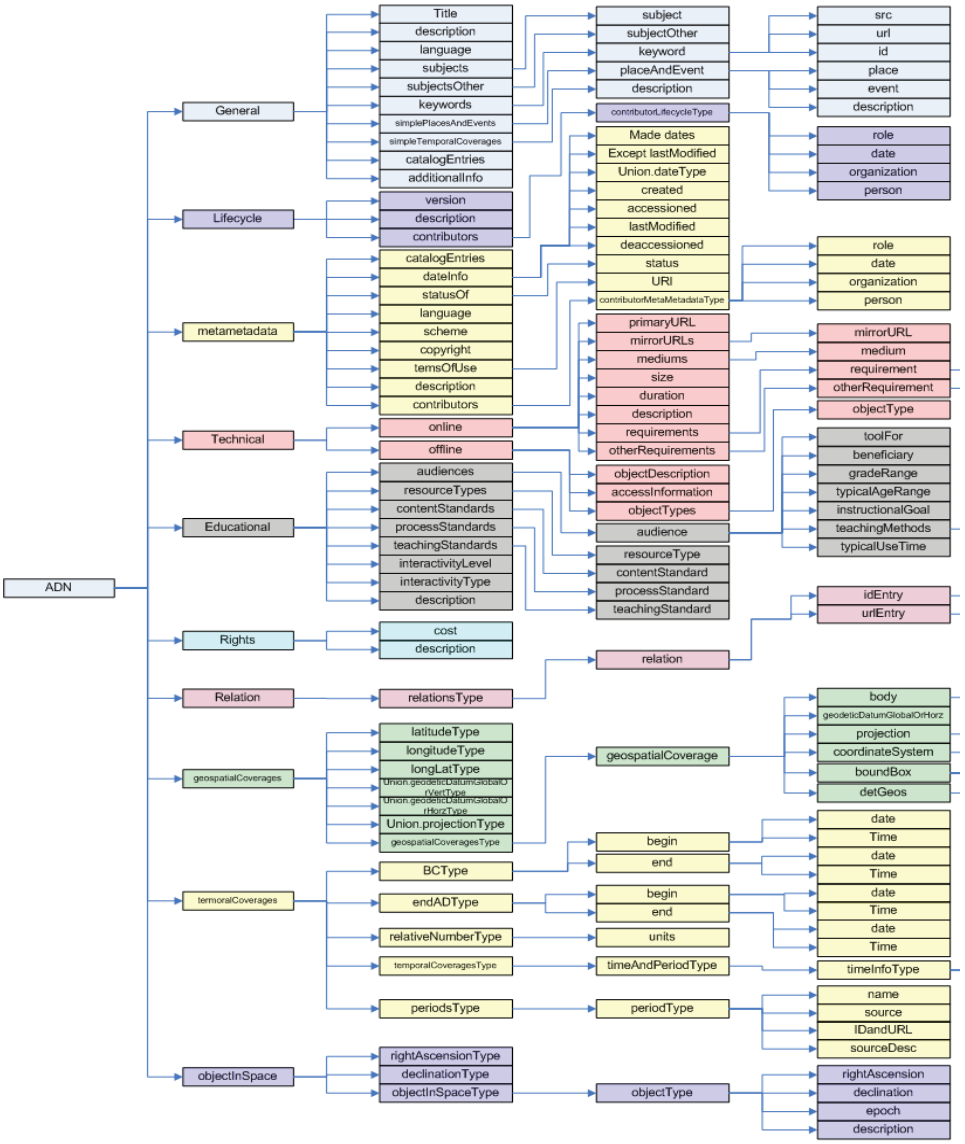


Figure 21: The ADN metadata schema

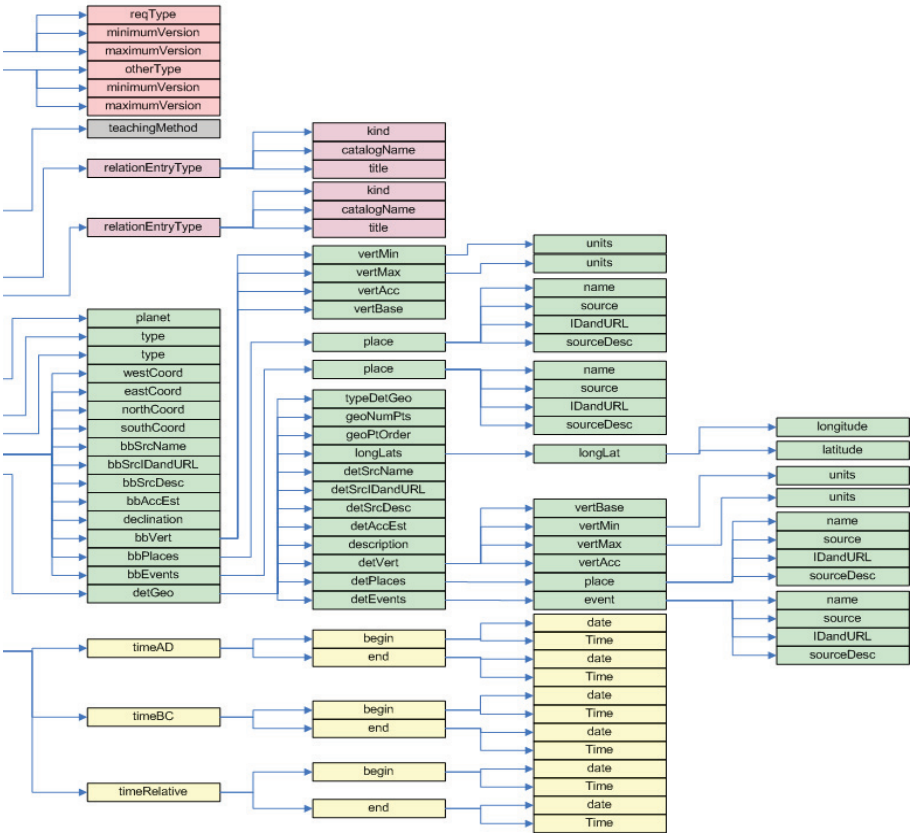


Figure 22: The ADN metadata schema (continued from Figure 21)

ADN Extension

ADN shares 7 of the 9 metadata sections in the IEEE LOM schema, excluding the sections "Annotation" and "Classification." These elements are not needed in the ADN schema (this will be explained shortly). The structure of the ADN standard is influenced by the IFLA FRBR model when describing learning objects [76, 77]. IFLA developed this model because they felt that a fundamental re-examination of the bibliographic record was necessary, largely to balance potentially divergent views and to respond to meeting an increasingly broad range of user needs and expectations [18]. ADN has done this in order to gain an extensive range of sub- and sub-sub elements. This has allowed the provision of a finer-grained schema for describing learning objects. To assure that the available learning objects are in accordance with American legislation, a special central committee evaluates learning objects before they are made available to the public. The metadata are professionally created. It is not possible for individuals to include local learning objects or make local comments without going through the central committee. This is why there is no need for the "Annotation" and "Classification" elements: The publishers and document users are not allowed to make annotations, and there is no room for local adaptations. The ADN presents itself as a stricter schema that is highly adapted for the exclusive use of professionals for metadata creation on a national level. The extensive use of elements also makes use of the schema for those querying documents more challenging since it is difficult to specify and interpret this level of detail without professional knowledge of the subject.

The schema has been implemented and is used in the DWEL project [32, 38]. This project has shown that combining efforts from different document providers can result in an information service with properties that suit a larger user group, in this case students and teachers at a K-12 school level. This has been done while ensuring that the learning objects is in accordance with legislation and with quality assurance. This project and the ADN are subject-specific. The ADN schema cannot be regarded as usable in a general educational context or outside of American framework. As a consequence, other American projects commonly employ the IEEE LOM schema instead of the ADN, including the efforts of the DC Educational Community [26].

SCORM

The Sharable Content Object Reference Model (SCORM) specification is a collection of technical standards and specifications for Web-based e-learning of the Advanced Distributed Learning (ADL) Initiative [21, 22].

The "Shareable Content Object" refers to how SCORM defines how to create LOs which can be reused in different systems and contexts. The "Reference Model" reflects upon the fact that SCORM is not a standard but a specification. SCORM is like an umbrella which uses a multiplex of existing industry standards. The SCORM specification references a range of existing standards and tells developers how to properly use them together in order to gain the desired compatibility. This includes works from the AICC, IMS, IEEE and ARIADNE. The purpose of this specification is to enable "plug-and-play" compatibility between e-learning systems so that they can

perform efficient sharing of LOs. By complying with the SCORM specification, this enables the LOs to be used at any e-learning system regardless of make or version as long as it is compliant with SCORM. This vastly extends the potential user group of each LO.

SCORM LOs can be anything from single files (such as a PowerPoint presentation), to complex educational systems based on packages consisting of multiple files (such as a web-page with separate images, video and sound) with interactive content. SCORM supports interactive, run-time communication, or data exchange, between the LO and the LMS. This enable e.g. questions and replies to be promoted through the LMS based on specifications stated in the LO. This enables interactive user experiences. Combined, this enables a full range of LOs to be shared.

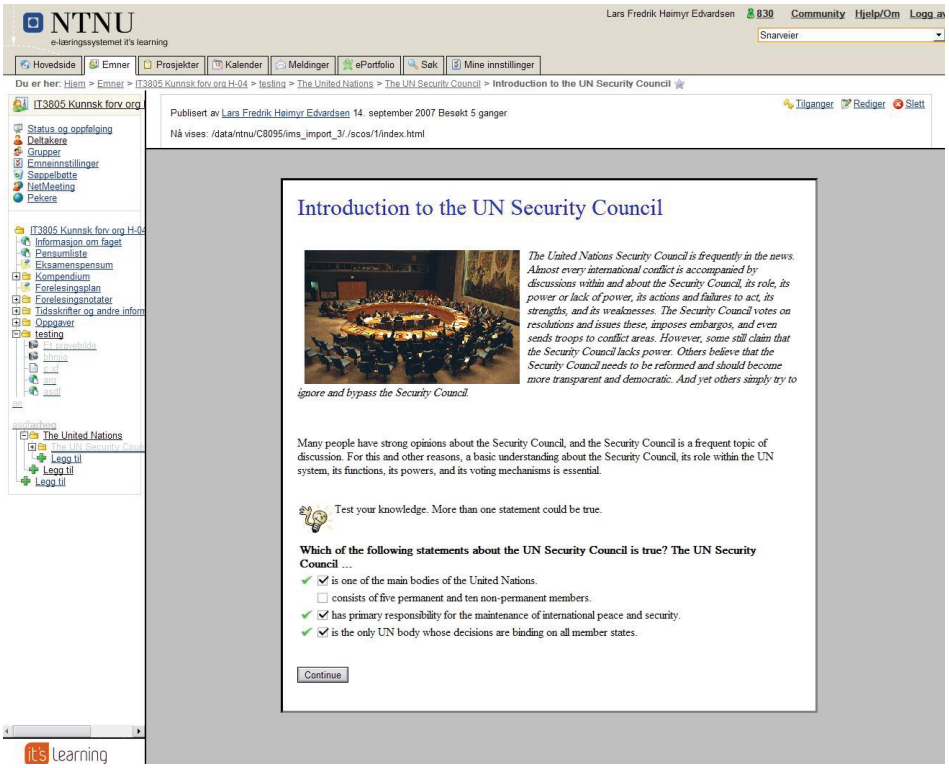


Figure 23: A SCORM LO imported into the NTNU LMS

Summary

For metadata in general and for metadata exchange, there has been an extensive focus on the Dublin Core. For educational documents, current efforts are now concentrated

around the IEEE LOM standard. There are no compatibility issues between the Dublin Core and IEEE LOM schemas. This research has used the IEEE LOM schema as its starting point for describing documents and their metadata.

Many metadata schemas contain element names that are multiple words joined together as one element name, such as “Copyright and Other Restrictions” from the IEEE LOM’s “Rights” section. This is a way to avoid the problematic issue of spaces for digital record processing. However, it is not ideal for humans to read. This research has therefore decided to present the element names with spaces. To clearly define where the element name begins and ends, this research uses quotation marks around the whole element name. If the element is part of a metadata section, then the section name is included in the element name presented. The element described above is hence presented as “Rights. Copyright and Other Restrictions.”

In order to achieve the purpose of the specification, there are a number of different aspects which needs to be specified. SCORM specifies how the LOs should be packed, how the content should be executed and how it is recommended used.

A SCORM LO consists of a:

- ZIP compressed file with all the content of the LO and the description files which specifies how to use the SCORM LO – both from the learner’s perspective but also from the computer execution perspective.
- The “imsmanifest.xml”-file contains the information required by the LMS to import and launch content without human intervention. This file specifies which files that should be executed.
- The “imsmanifest.xml”-file is also used for submitting advanced metadata using the IEEE LOM standard.
- The actual LO(s).

SCORM LOs are extremely user friendly in terms of providing the educational audience extensive information regarding the LO’s usage and educational context. However, manual creation of these is not developer friendly. As a result, there is a lack of SCORM object even though the LOs are available.

SCORM does not create educational resources for you. There is hence a need for manual efforts to specify what to present, and how the presented material should be used by the LMS and by the user. Nearly all LOs published at the NTNU LMS are not complex educational systems but rather individual files with the LMS and user actions pre-specified. Here the LMS is used as a distribution channel with the possible user interactivity included as part of the LO. E.g. PowerPoint presentations with interactive content. This standardized property of the LOs enable usage of default content in the “imsmanifest.xml”-file to specify LMS and user behaviour. There is, however, a need for LO specific IEEE LOM metadata to promote the actual LO content and the context in which the LO has been promoted used at.

2.4.4 Learning Objects and creation of Learning Object Metadata

Koutsomitropoulos et al. expresses how the IEEE LOM standard can be a valuable tool for describing learning objects as a tool in order to help students find the most desirable learning objects available on record [89]. They address the issue of regarding complexity and cross-metadata language compatibility in educational metadata standards, finding the need for ontology in order to generate mappings between the IEEE LOM and DC-ed. This project collected metadata from an existing metadata repository and hence did not have to address the issue of generating the initial metadata.

Di Nitto et al. addresses the issue that current Learning Object metadata standards are young and still needs to develop in order to gain the needed elements for describing education resources [31]. However, the introduction of more elements will increase the complexity of the metadata standards, and increase the amount of time it takes to describe each Learning Object with a complete metadata record.

The JISC MOSAIC project expresses the benefits of having extensive metadata descriptions. Some AMG efforts are imposed, such as extraction of technical metadata, such as file format, size, creation date, from files [35]. Automatic generation of non-technical metadata and addressing metadata quality issues has yet to be presented.

The Metaspeed project addresses the importance of metadata:

> *In today competitive business environment the proper management of organizational digital resources is crucial for making timely decisions and responding to changing business conditions. ... However digital resources are increasingly being recognized as a very important organizational asset au par with finance and human resources.*
>
> Peneva et al. [117]

The Metaspeed project is working to address issues regarding cross standard compatibility between metadata standards and automatic metadata generation [117]. A detailed task description and project results have yet to be published besides their efforts in working with SCORM and MPEG-7.

A primary concern within the Learning Object community should be the availability of Learning Objects, or the lack of such. The amount of Learning Objects available for download on the internet seems to be on level with the amounts we experienced a few years back, even though the amount of documents on the internet has extended significantly. Meyer et al. propose usage of Wikipedia articles as basis for generating Learning Objects [103]. They used a Machine Learning approach for generating classification.

Motelet et al. propose usage of a Graphical User Interface (GUI) where standard or common entities of IEEE LOM elements are promoted to users [105]. This is done in order to speed up the metadata creation process. Once the metadata has been created,

the system generates a finished Learning Object based on the IEEE LOM specifications, similar to the SCORM creation process presented in Article P5 (see p. 239).

2.4.5 It's learning - The NTNU LMS Intranet

The use of LOs in Norway is increasing on all educational levels [87]. At NTNU, the current LMS lets its more than 30 000 users create LOs based on predefined templates. These templates contain predefined property sections where the user fills in content in order to create the LO. There are templates available for creating exercises, tests and inquiries, as just a few examples. In addition, users can upload stand-alone documents from their personal computers to be included as part of the LMS's LOs. The documents submitted are very diverse, both regarding their educational content and visual appearance. The span in diversity ranges from documents based on predefined official administrative templates created by university employees, to documents without any apparent structure created by students on private computers.

This LMS is not made for LO sharing, and hence sharing of LOs is not allowed. As a result, large collections of valuable educational documents are each year locked away and hence made unavailable for the vast majority of potential users. Currently only the handful of people who attended the specific course in the specific year and semester in which the LOs were published have access to the LO. This set-up also lacks the ability to make efficient queries and is thus a vast waste of valuable documents, which has direct influence over the LMS users' practices:

- Users are not able to locate their own previously published content.
- Users cannot share LOs with other students, lecturers and guests without having to perform a work-around or creating new a LMS user.
- Users cannot use the LMS to locate existing documents made by others.
- Documents must be continuously recreated, because it is not possible to build on existing documents.
- Research is curtailed or limited because documents cannot be shared, which discourages cooperation between educational disciplines.

A new LMS or a new LMS version is being evaluated in order to enable the sharing of LOs at NTNU and between other universities and university colleges in Norway and throughout the world. This would enable a large number of document authors and document users to share documents, knowledge and experience across larger physical differences and between technical and organizational boundaries. The targeting of such a large user group, which does not share the same user background, requires more descriptive information about document content in addition to existing identifiers. A proposal has been made to use an international metadata schema with extensive LO description possibilities, the IEEE LOM [74] or the Norwegian version, which is NORLOM [46]. These schemas include more than 60 unique descriptive elements. Generating entities for these elements will to a large extent have to be performed by AMG algorithms. AMG can be used to collect or create metadata by:

- Harvesting embedded metadata from existing documents.
- Generating new metadata from a document based on an analysis of the individual document.
- Generate metadata based on the publishing context of the document.
- Inheriting metadata from existing, prior versions of the document.

The majority of documents in the case LMS are published once, limiting the value of using the document's heritage as a source for a primary AMG method. The context descriptions can provide valuable information regarding collections of documents, but they do not describe the specific document particularly well. AMG methods will therefore have to be based on the document itself, either by harvesting or extraction. Harvesting metadata can be used to generate a number of elements, by collecting the metadata created by the user and the user's content creation software. However, such elements should only be used if they actually reflect the specified document. False entities should be avoided. The diversity in published documents, in regards to document formats, visual appearances and the multi-lingual environment, reduces the effectiveness of existing AMG extraction efforts. To tackle the challenge of generating correct metadata, any AMG effort needs to be based on a more reliable data source.

3. Context and Research Design

The following Chapter gives a more in-depth description of the context of this thesis. Firstly, Chapter 3.1 gives more insight into why reaching the research goals are of importance. Secondly, Chapter 3.2 presents the process in which this research has been conducted.

3.1 Reaching the Research Goal

As described in Chapter 1.3, this thesis started with the challenge of metadata in an educational context. However, it soon shifted towards looking at why people "don't create metadata" and how metadata could be created by computers as a supplement to manual metadata creation efforts.

The most basic form for creating metadata without human interference is to retrieve metadata that already exist within the documents. This lead to the first research question:

RQ1.1: What is the quality of automatically generated document content (embedded metadata and document formatting)?

Nearly all end-user applications automatically generate metadata. All common Word processors, spread sheet applications, presentation applications and image processing applications generate vast amounts of metadata. In addition, vast amounts of metadata could be inherited from previous versions of the document if a document is based on a template. A vast amount of metadata is of little value if the elements registered are not relevant to you. In addition, the quality of the relevant metadata is vital in order to provide value to the user of the metadata.

As the State-of-the-Art presented, related AMG research and commercial products have been heavily focused on one document type with similarly formatted documents. In this thesis' view this does not reflect upon real-world scenarios particularly well. People in general are notoriously known for being less structured, having a hard time of sticking to technical specifications and of having creative will to do "things" in their own way. This is a vision of the document authors that the AMG algorithms should be designed around as a basis for their efforts. Without such flexibility to expect the unexpected in the published documents, the usefulness of AMG in a general context vanishes. Or to put it in another way: AMG algorithms created in one context seldom generate valid metadata if they are moved to a new context or if a different type of document is submitted. If the document type changes, then the AMG algorithm need to be changed as well. If a new user type is to use the AMG algorithm, they cannot do this, since it can't handle multiple data sets. You cannot include documents created in another department since you haven't used the document template.

Such restrictions make extensive restrictions to the usefulness of existing AMG algorithms. The AMG algorithms must be constructed to expect the unexpected. And regardless of the document content be able to locate the specific content that is essential in order to identify the document and to create relevant metadata. The AMG algorithms

must handle each document individually, maybe even combined in order to generate the best results. This led this thesis to the second research question:

RQ1.2: Can AMG approaches be combined or selectively used on a document-by-document basis?

If such functionality could be achieved, then this thesis would pave the way for fare increased usability of AMG algorithms in document collections where there is little or less document structure – non-homogeneous document collections. This movement away from document collections based research documents and library collections would be significant in order to introduce active usage of AMG in contexts such as company intranets, on MS SharePoint sites and in personal document archives.

To do this, an analysis of the actual document file content, the so-called "document code", is central to learn about the content of each document. This thesis needed to find common lowest common denominators among all documents, regardless of the document's visible content. Through analysis of the document code of common file types, such lower common denominators were located, enabling identification of the document's intellectual contents created by its users rather than template contents and contents of questionable quality created by content creation software (and user applications such as MS Word, PowerPoint etc.). Basing AMG efforts around the document code can enable detailed, structured and correct metadata from non-homogeneous documents.

At this stage this thesis knew how AMG could assist in creation of high quality metadata that describes each and every document in the dataset. Still, in the educational context there are few documents, or Learning Objects, that are shared as metadata only. LOs are commonly shared as a package consisting of both metadata and the LO. So, could AMG assist in creation of such packages of metadata and LOs?

RQ1.3: Can AMG enable automatic generation of complex sets of metadata, enabling usage of advanced Learning Object document formats, such as SCORM?

Of the various LO document formats, SCORM is a particularly exiting format as such objects consists of the original LO plus a metadata file packed within a single ZIP-file. This is of essence, as the original LO remains unchanged and the original applications still can be used, while educational LO metadata are present.

All of these research questions contribute in order to address the main research question:

RQ1: Find methods to automatically generate metadata from non-homogeneous document collections for promotion of educational resources.

This research question is extensively addressed throughout this thesis. This thesis has shown a range of candidate data sources that can be used for AMG efforts, how various AMG efforts can be executed based on conditions on a document-to-document basis in order to achieve the highest possible data quality, and to generate the final LO that the user strives to distribute and enable efficient retrieval.

Combined these methods and techniques achieve this thesis' motivation: To enable more efficient sharing of knowledge through distribution of LOs that contain extensive and high quality metadata to maximize the LOs potential of being located and reused. And doing this without placing technical or metadata knowledge requirements on the LO author or publisher.

3.2 Research Process

This thesis started by studying metadata and metadata schemas. With highly advanced and detailed metadata schemas available for describing LOs, this thesis wanted to grab hold to existing LOs and their descriptions. But there were hardly any LOs and associated educational metadata to retrieve. Further analysis reviled that only a few hundred LOs with associated educational metadata had been created *worldwide*. This while there were *millions* of LOs being published without associated educational metadata. Sharing of educational resources was clearly not helped by the presence of educational metadata. Or maybe that was a reason why there are so many similar LOs on the Internet – Because existing LOs are not efficiently reused since other teachers do not know of these and hence need to create their own LOs.

When looking at the local LMS at NTNU, It's Learning, the world view did not become significantly more positive. Several small-scale qualitative studies were conducted of It's Learning where published LOs were analysed in terms of presence of metadata and the quality (correctness) of these metadata. The analysis also showed that It's Learning had support for importing LOs with educational metadata descriptions in the SCORM format. On documents were located that were based on SCORM.

This way of looking at sharing of LOs is not very positive. But it became a motivation for creating the first paper: "Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment" [REF].

In the hunt for the lowest common denominators among the shared documents from It's Learning, the contents of the document code started to take the stage. By decomposing the documents into their document code, contents created by the users, inherited from the document templates and included by end-user applications became visible. But often not labelled. Extensive research were conducted in order to determine what characterized content created by the users, inherited content and by applications. This is a major challenge reflecting upon the resulting data quality, as contents can be created by one, two or all three actors in various documents.

A side effect of basing AMG efforts on the document code is presented firstly in paper P2. This is the ability for the AMG algorithm to *not* generate entities. If the desired content is not located, the AMG algorithm should return no result. Other State-of-the-Art AMG algorithms consistently create entities regardless of the content of the document, resulting in a lot of false data when conducted on a non-homogeneous document collection. This ability to be restrictive to creation of entities is very significant, as it enables AMG efforts based on the document code to be the first AMG effort in a sequence of efforts, guiding document contents to further processing in order to achieve high quality metadata.

In the hunger for more educational metadata, all possible data sources were evaluated including

1. The file it selves based on
    a. The document code
    b. The documents' visual appearance
    c. Intellectual content
    d. Non-visible formatting
2. Contextual information from the publishing site
3. User profile information from the publisher
4. Connections to third-party systems for extending the publisher information
5. Registration of actual usage of the documents based on information of
    a. Who downloaded the Los (including their user information) and
    b. Statistical information of use

Of the list above, the two first sources became the main basis for future data sets. Information regarding who downloaded each file, connections to other systems and statistical information were either not obtainable due to privacy concerns or were simply not registered in the system logs. Several sequences of AMG efforts were created in order to determine an optimal way of conducting AMG efforts.

A real challenge still remained though; How to evaluate what was a good and what was a bad result. In other words, the scale to use in order to determine the quality of the generated entities. Luckily the thesis supervisors were familiar with the works of Lindland et al., which in time became the framework for this thesis to determine the quality of entities.

The initial quantitative results were published in paper P3, while paper P2 performs more in-depth analysis in a quantitative study based on papers from the same data set.

This thesis gained access to a vast amount of content on the NTNU LMS It's Learning. The SP1-paper was created in order to review and document how It's Learning was used for educational purposes at NTNU.

Similarly, this thesis gained access to a second dataset from an Auditing firm in order to compare the quality of automatically generated metadata. In addition, the auditing firm was used to illustrate how to use document templates to promote desired usage. This thesis gained experience in making changes to corporate templates in order to enable more efficient document retrieval. This by promoting the document content which were of special importance for the Auditors, without changing the visual appearance of their familiar documents. The paper SP2 were created, but it was not published.

Metadata is of limited value if the metadata is not being used. Based on knowledge of how high quality metadata can be automatically generated, this thesis explored how AMG could assist in various situations. Firstly, a study was conducted for automatically generating metadata based on published papers. This study was conducted in order to promote that AMG can be used, even though the data set is virtually homogeneous

(very visually similarly formatted documents all with intellectual content in the same language). This study was published in paper P4.

Next, the constraint of homogeneous documents was once again lifted, and focus on the educational was enforced. How could we promote sharing of educational resources or LOs? This thesis knew how to use almost random documents to create educational metadata. And of the educational data formats in practical use, SCORM were among the most commonly used. It was also the only LO "package" of LO and metadata that was supported by It's Learning.

In paper P5 this thesis shows practical usage of AMG to generate LOs consisting of an educational resource and extensive educational metadata descriptions, all created with an absolute minimum of requirements placed on the human user of the application. These both in terms of knowledge requirements and in terms of time needed to create metadata.

The various threads and angles created in the various papers were merged and put into a common perspective in paper P6.

In sum this thesis has had a practical focus for metadata and educational resources. Standards and technologies are of limited value if they are not used. This thesis has shown how new and existing technologies can work together in common benefit in order to promote sharing of educational resources.

4. Research results

This chapter dives into the research results. First up is Chapter 4.1 which explores how the environment in which the document is created influences the resulting document. Here we find a clear distinction between documents created in the system controlled environment and stand-alone documents. The documents created in a system controlled environment show a unique consistency: All the documents are created from a small number of pre-defined templates, there can be enforcement of mandatory sections, and though usage of log-in features the system controlled environment can be certain who the document author is. However, these characteristics do not ensure that the desired metadata quality is achieved, only that the created content shares a multitude of characteristic.

Stand-alone documents can be created in an infinite number of ways. Still, statistically people use the same applications to create their documents. Due to this, there are considerable similarities between most documents regardless of their visible appearance or intellectual content. This chapter presents how the source code of documents ("the document code") can reveal hidden structures and how converting between document formats might corrupt visible and non-visible contents from documents.

Chapter 4.2 explores quantitative characteristics from stand-alone documents retrieved from NTNU's intranet. Here we find a combination of documents created in a system controlled environment and in a user controlled environment. The stand-alone documents were all initially created in a user controlled environment, but were shared in a system controlled environment which enforced the user to act in specific ways in order to be allowed to publish the document. For this analysis, this research gained access to 424 published LOs, of which there were 289 stand-alone documents. As these documents were gathered during the pre-study phase of this thesis, these documents are referred to as the "pre-study dataset".

The majority of the uploaded documents are in Adobe PDF, MS PowerPoint or MS Word document formats. Virtually no documents contained either an educational metadata description or an informative description. It is evident that the document authors and publishers use minimal efforts in giving the document a semantic metadata description. So few in fact have given a semantic description besides the Title element, that it is highly questionable if the documents authors and publishers are aware that such content can be stored as part of the document.

Quite worrying is the fact that a number of entities were misleading or directly false. Some elements even had multiple, conflicting entities registered. This reflects on both semantically and technical elements which are not user specified. There is considerable uncertainty regarding the quality of the gathered document metadata and regarding the awareness to metadata by document authors and publishers.

Chapter 4.3 explores the details of selected elements by performing a qualitative analysis. For this analysis, this research gained access to about 11% of the courses at NTNU and in total 3483 stand-alone documents published from these courses. These

documents are referred to as “the final dataset”. These efforts of the analysis are threefold:

Firstly, the impressions from the quantitative analysis are confirmed: Embedded metadata from documents contain a high latent possibility for false entities. Without any central control over a document, the document’s content is highly influenced by the local applications used to create it. These analyses have confirmed that automatically generated entities generated by document applications frequently contain false entities.

Secondly, the effects of existing AMG harvesting and extraction algorithms are explored as the algorithms show just how limited most existing AMG efforts are when being executed on a diverse document collection. Most of these AMG algorithms generated entities regardless of document content which resulted in a vast number of false or partly false entities. Or when combined with other AMG efforts; a high number of candidate entities.

Thirdly, we look at this thesis’ new approach of using the document code as basis for AMG efforts. This section shows how the document code can be used to guide the right AMG efforts to their optimal content while avoiding known content of lower quality. Hence, we can explore usage of multiple previously developed AMG task- and subject specific logics as part of the same AMG rule set, as the AMG efforts based on the document code guides all the other AMG efforts to their optimal data source. And is vast contrast to previous AMG efforts, if the optimal data source is not found, then the AMG efforts are not executed. This type of logical selection and prioritizing document sections has previously not been possible to achieve on a document collection like this with extremely diverse documents.

Though, first up is the process of creating educational documents.

4.1 The process of creating educational documents

In theory every document can be created in its own unique way. In practice there are extensive similarities in the actions taken by users and software to create a document, even though the intellectual content of the document itself may be unique. This chapter presents how user environments affect resulting documents in regards to semantics, consistent properties, content validation and embedded document metadata. This initial analysis is based on 3600 LMS documents and their attached stand-alone files collected from 55 different courses. The content of a document is strongly coloured by the environment in which it was created. The fully controlled and the not-controlled environments are the two extremes in user environments. Here the fully system controlled environment represents a system without any opportunities for individual adaptations of the documents, while the not-system controlled environment enables the user to make all decisions regarding use of application(s), document templates and document content.

Chapter 4.1.1 presents more detail about the differences between these user environments.

Chapter 4.1.2 presents more detail about how documents are created in a system controlled environment, by analysing the case LMS and how documents are created in this specific system controlled environment and the effect this has on the resulting document code.

Chapter 4.1.3 presents how stand-alone documents are created. This analysis uses the creation of Word documents as an example of a document format. This chapter continues by analysing the influence of a conversion process on the document code when converting a Word document to the non-compatible document format of a PDF document.

4.1.1 Different document creation user environments

The case LMS is a representative of a system controlled environment with fully controlled features. The LMS is accessible to the user by logging in over a network connection. The LMS only allows a specific application to be used to create documents, though the user is allowed to choose the pre-specified document templates upon which to base the new document. The LMS conducts content validation of specific document content, such as presence of a document title and validation of dates. The available templates are task-specific with a limited set of pre-specified document properties. The user needs to select the correct template in order to obtain the task-specific properties of the document type. Such document templates make users aware of specific properties of the document and encourage them to describe the document in a standardized way with content that is visually present [36]. This ensures the creation of documents with identical syntactic structure. The templates are used nearly exclusively in accordance with their intended use because other possibilities are restricted. These properties give each document a structure and consistent properties that are present in all of the documents created based on the LMS document templates. In addition, the LMS provides context information describing the user, the publishing system, storage section information, and other context information regarding the section published. The LMS thereby provides AMG algorithms with multiple data sources containing systematic and consistent properties. Chapter 4.1.2 presents more detail regarding the properties of the system controlled environment, the properties of the document templates and the content of the finished documents.

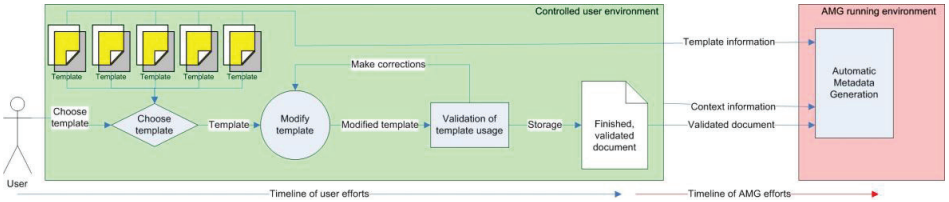


Figure 24: Documents from a system controlled environment as data source for AMG efforts

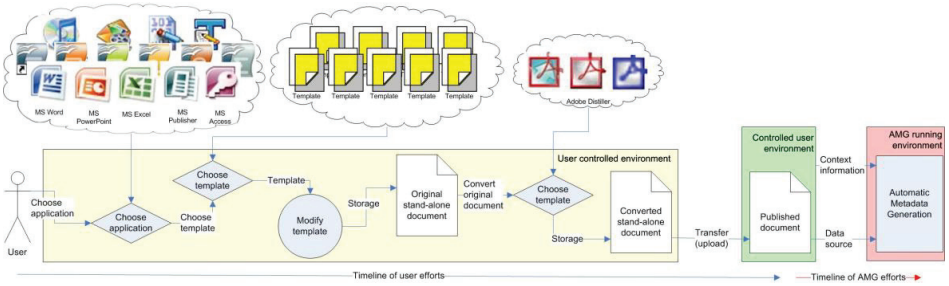


Figure 25: Converted stand-alone documents as a data source for AMG

The restricted properties of LMS documents make stand-alone documents a popular alternative: close to 75% of all publications contain one or more uploaded stand-alone documents as an attachment to the LMS document. Stand-alone documents are created in a not-system controlled environment. This gives the user the freedom to choose application(s) and templates, and the freedom to choose how to use these resources. These qualities give the user extensive freedom of expression at the expense of systematic and consistent document properties. Stand-alone documents are frequently converted before being published, e.g. from Word to PDF document formats. This affects their content:

- Content can be added, altered or removed; non-visible formatting data is commonly discarded.
- The converted document can contain metadata that reflect the converted document but not the original.
- Documents can be subject to security restrictions, which prevent AMG algorithms from accessing their content.

Additional uncertainties regarding converted documents increase the vulnerability of metadata harvesting to generate false metadata. The LMS shows extensive varieties in regards to published stand-alone documents, as all such documents are accepted for publication. This research found 41 document formats, a range in content types (texts, spread-sheets, presentations, etc.), content qualities (from informal notes to papers) and intellectual content in a multitude of languages. The stand-alone documents have a diverse visual appearance, ranging from being based on predefined official administrative templates used by university employees, to documents without structure created by students on private computers. The structured properties and consistencies found in the LMS documents are hence not found in stand-alone documents. Chapter 4.1.3 presents in more detail the properties that characterize these stand-alone documents.

4.1.2 Creating documents in the system controlled environment

Basing documents on pre-defined templates

Learning Management Systems (LMSs) are commonly used to provide additional services that stand-alone documents cannot provide or to provide document types that are easy to create and administrate. Such systems usually enable sharing of educational content in a standardized way and where the user's technical barrier for creating publications is low. This enables a larger user group to employ the system without having to undergo extensive training. Most LMSs are system controlled environments. This means that users need to follow pre-ordained rules to use the system. These are requirements set by the system provider or system administrators.

The case LMS uses document templates to enable users to create desired document types. Document templates make the user aware of the document's specific properties and encourage the user to describe the document with visually present content in a standardized way [36]. This allows more users to create documents with the desired properties. In the system controlled environment, the user is guided and forced to comply with the opportunities and restrictions that are provided by document templates and enforced by the content creation software.

The case LMS has restricted publishing possibilities based on the user profile: The user must log in to the LMS before he or she can publish a document. Publication can only take place in sections where the user is allowed access, meaning specific courses.

The process of creating documents based on templates

In a system controlled environment, the user is only allowed to create documents based on existing document templates. The user is not allowed to create his or her own templates. Instead there is a third party who is the only one allowed to create templates. These templates are different from templates for stand-alone documents, in that they enable use of administrative tools that are provided through the LMS. Common central administrative tools include user group access control, time restrictions regarding document availability and management of student deliveries.

These templates have pre-defined sections intended for specific content described in a schema. The restrictiveness of the document schemas is used to encourage the user to use the template in accordance with the system's schema. This is typically done by providing special template-specific visual characteristics in the resulting document or special administrative tools for that specific document type, such as administration of delivery dates. By complying with the system's schema, the user then has something to gain that cannot be obtained by using other document types.

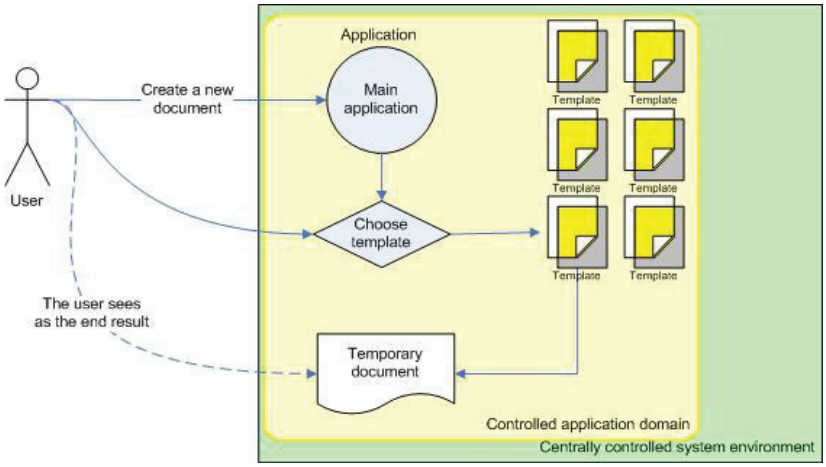


Figure 26: Creating a new document in a system controlled environment (stage 1)

When a new document is to be created, the user is faced with the choice of deciding which template to use for the document. Based on that decision, the user is presented with a specific template with template-specific properties and possibilities. In the case LMS, the user can create documents based on these specific template types:

- **File** ("Fil"): Used for uploading stand-alone documents to the LMS.
- **Link** ("Referanse"): Consists of a single hyperlink.
- **Note** ("Notat"): An undefined template consisting of a single text section.
- **Exercise** ("Oppgave"): Can consist of an exercise text, multiple uploaded stand-alone documents, with exercise delivery possibilities and grading and correction possibilities.
- **Image with description** ("Bilde med beskrivelse"): Consists of an uploaded image and a description.
- **Process oriented document** ("Prosessorientert dokument"): A document type that is adapted to the users' actions though multiple sub-steps.
- **Explanatory sequence** ("Forklaringssekvens"): A sequence of steps designed to explain a concept step-by-step.
- **Test** ("Test"): This is an online test that can contain the test, give the test to students, allow instructors to grade the test, and present the results to students.
- **Inquiry** ("Undersøkelse"): An inquiry where the interviewees answers questions.

In addition to documents, the case LMS allows users to customize their own course-specific section of the LMS by creating folders ("mappe") in which documents can be kept.

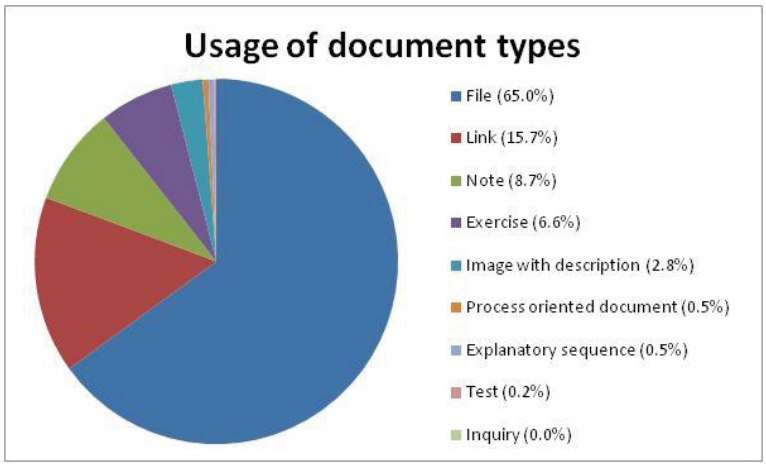


Figure 27: Percentage of use of document types in the case LMS

Figure 27 shows use of the case LMS's document types. These observations are in line with other analyses of the LMS and its usage [88]. Use of the case LMS document types was recorded during the pre-study phase of this research. Analysis of the LMS content from 55 courses showed that two-thirds of documents created in the case LMS were of the "File" type. Only 12.8% of these documents contained a content description. Instead, the "Title" element was often used to give a short description for identification based on the educational context given through other published objects and lectured content. Close to 75% of the published documents have the ability to include one or more uploaded files. Of the other available document types, "Link" is frequently used to publish hyperlinks, while "Note" is used to present all types of plain-text content. Document types and document content that were not intended for public display have not been collected for this research.

Template type content

In a system controlled environment, templates can be used to manage the content specified by the user, because the application can enforce compatibility with the given template schema. The user is only allowed to submit content for the document through the pre-defied sections of the template. These sections are commonly named and presented to the user to indicate what type of content should be included in the specific section. These template sections can be governed by the publishing system. Enforcing this functionality ensures that the document vocabulary used complies with the template schema, thus avoiding conflicts with controlled vocabularies [141]. This means that the application can enforce the use of mandatory template sections and validate restricted value spaces. If the schema requirements are not met, then the application can refuse to store of the document. If the applications do not enforce the schema requirements, then it is up to the user to ensure that the schema requirements are met.

Templates can be presented as a dialog where the user supplies content to the fields that are presented. This enables dynamic, multi-stage templates to actively guide the user

through the document creation process in smaller and easier steps than creating the entire document with all its properties at once. Such applications are commonly referred to as "wizards." These must be adapted to the mental model of the user group's understanding of what the application should do [146].

In addition to the user-specified document content, there is the possibility of including centrally administered context descriptions:

- Firstly, descriptions of the technical placement of the new document must be recorded, such as placement in a specific subject, within a folder, subfolders and so on.
- Secondly, descriptions of the subject's context, in which specific elements can be collected from a centrally administrated course profile, can be included. The LMS can in turn base its course profile on a course profile retrievable from another centrally controlled computer system.
- Thirdly, if the user logs in, then user information can be included: These user profiles can contain full name of the user, the user's role in the course or possibly a complete vCard. The user profile can in turn be based on harvested data from a centrally controlled user registry.
- Fourth, the LMS can base its timer on a centrally controlled clock. The time of creation and modification is then not affected by local time variations that can occur as a result of differences between users' personal computers.

In the case LMS, all published documents are automatically labelled with administrative data and data specific to the document type. The administrative data includes the publisher description, published date, placement data (course, semester and folder(s)). For each document template type, there is a template-specific creation tool that displays available document content elements and enforces compliance with mandatory schema regulations (value spaces). These elements can be seen in Table 6. It is mandatory for all document types to have a title, which has to be provided manually. Aside from this element, the user decides how to use the remaining elements.

Selecting the right document type gives the publisher the ability to specify valid administrative document properties. For each of these administrative properties there is functionality within the LMS that administers the usage of the document in accordance with the described elements. Specifying the document type and its properties has a direct influence over the potential usage of the document. For example, the template type "Exercise" in Figure 29 enables the use of administrative properties for enforcing delivery dates and grading of student exercises. These are functions that are not available for other document types.

Restricted value spaces are displayed as Boolean alternatives or as a pull-down list when the document is created. It is not possible to specify entities other than the ones listed. In the creation of the document, these elements are assigned a default value. This value is a valid entity. Because of this, it is not possible to distinguish between elements that are not used by the publisher and elements which were given the correct entities by

using the default value. It is therefore not possible to evaluate the degree of actual usage of these elements without questioning the publishers.

The system controlled environment does not enforce correct usage of all template sections. The document type "Link," presented in Figure 30, allows the creation of hyperlinks to content outside of the LMS. The LMS does not validate if the user-specified URL complies with the schema definition for valid content of its "URL" element. It is therefore not certain that the entity complies with the LMS' schema. Because of this, the entity of the "URL" element cannot be fully trusted to be valid, even when it is deliberately specified by the user

Table 6: LMS document types

The LMS document types	LMS type	Course name	Course semester	Publisher name	Published date	Title (one text line)	Description (multiple text lines)[6]	Reference (one text line)	Stand-alone document(s) as sub-element	Mandatory (Boolean)	Delivery deadline (date & time)	Delivery administration[7]	Acceptable tries (number)	Anonymous delivery (Boolean)	File name	File size
File	A	A	A	A	A	M	M		U						A	A
Note	A	A	A	A	A	M	M									
Link	A	A	A	A	A	M		M								
Image with description	A	A	A	A	A	M	M		U							
Exercise	A	A	A	A	A	M	M		U	M	M	M		M	A	A
Process oriented document	A	A	A	A	A	M	M			M	M	M				
Test	A	A	A	A	A	M	M			M	M	M	M	M		
Inquiry	A	A	A	A	A	M	M			M	M	M				
Explanatory sequence	A	A	A	A	A	M	M									

(A = Automatically created, M = Manually creatable, U = Uploadable)

[6] This element has multiple synonyms depending upon the LO type. For example, the LO type "File" calls it "Comment" ("kommentar"), while the LO types "Note" and "Image with description" call it "Text" ("Tekst").

[7] Available for lecturers. Displays delivery information that includes who has delivered their assignment, the delivery time (day, hour, minute) and the delivery as a file (LO type "Exercise") or online answers (LO types "Process oriented document," "Test" and "Inquiry").

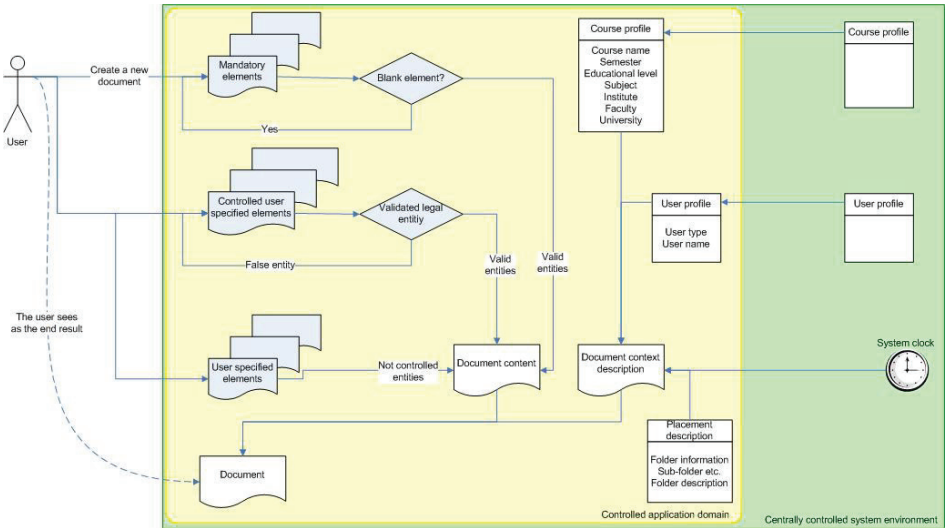


Figure 28: Creating a new document in a system controlled environment (stage 2)

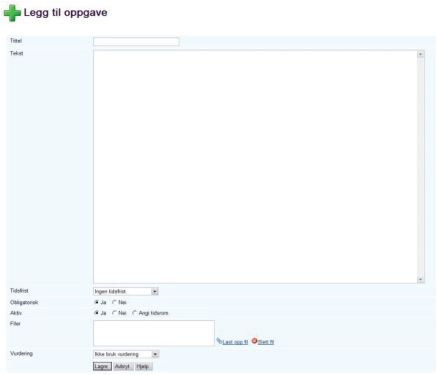


Figure 29: It's learning template for Exercise document

Figure 30: It's learning template for Link document

Templates can pollute data as a result of template content and default values, such as when several of the case LMS' document types are given default document properties when created. Elements such as "Mandatory" (for exercises) are set to "Yes" at default. Properties of the document can therefore reflect other interests than those of the user.

Uploading stand-alone documents into the system controlled environment

Users can have many reasons for wanting to upload stand-alone documents instead of creating documents based on LMS document types. The most common reasons are:

- To enable usage of application functionality that is not supported by the system controlled environment, such as spell-checker, document merging facilities and increased formatting possibilities.
- To allow distribution of existing documents, such as pre-made exams or print-outs or articles and lecture slides.

Stand-alone documents cannot be imported into the case LMS as a document type. Instead, stand-alone documents can be uploaded as part of a LMS specific document type. These documents need to follow the schema regulations as do all other document types, although the content of each uploaded stand-alone document is not analysed by the LMS. Rather, stand-alone documents are commonly included as an attachment to the system specific document type. The stand-alone documents can therefore keep their original properties.

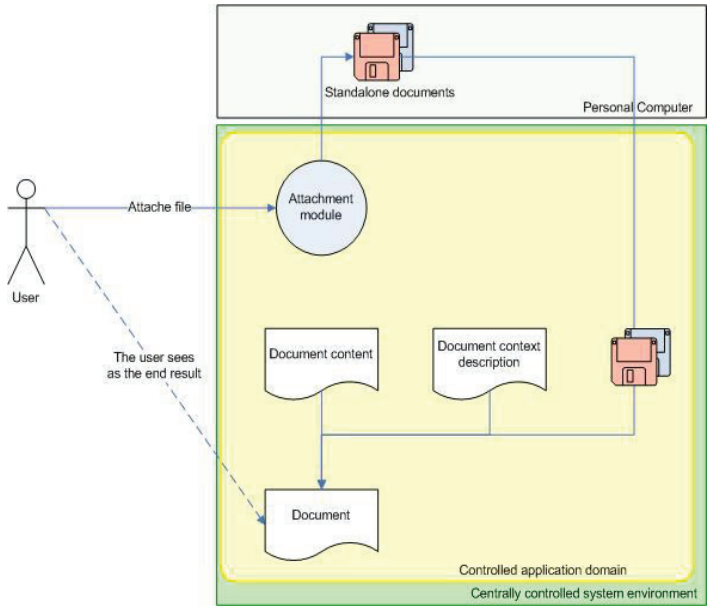


Figure 31: Uploading stand-alone documents to an existing system document

4.1.3 Creating documents in a user controlled environment

Document templates are the basis for creating stand-alone documents in the user controlled environment as well as the system controlled environment. The distinction

between the two environments is their use of content creation software, the number of available document templates and enforcement of the template.

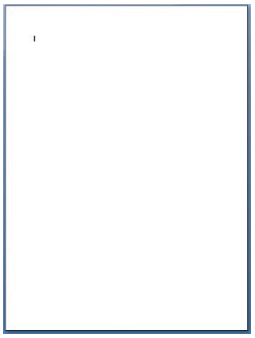


Figure 32: Blank Word template

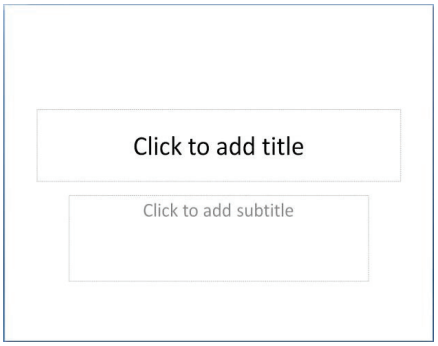


Figure 33: Blank PowerPoint template

Figure 34: NTNU lecture slide PowerPoint template

Figure 35: NTNU thesis PowerPoint template

In the user controlled environment it is up to the individual user to decide which content creation software to use and how to use these tools in order to generate the document, including its metadata, formatting and intellectual content. The templates can include or be without visual content. Figure 32 shows the MS Word default template "blank.dot," while Figure 33 shows the MS PowerPoint default template "blank.pot," which contains visual content. Organizations can use templates to create a common identity and to standardize the appearance of official documents, as in the templates in Figure 34 and Figure 35.

The content of templates can be a disadvantage in regards to AMG if undesired or unintended content in the template can be inherited by documents that use the template. For example, several of NTNU's stand-alone document templates contain elements with pre-defined entities:

- Creator = “O. Rakel”
- Title = “Line one”

If the document’s elements are not updated with valid entities, then the resulting document will contain false metadata that reflects the template and not the resulting document.

Creating a new stand-alone document

To illustrate the processes involved with the creation of a new, stand-alone document, this research presents the creation of a Word document. The creation of Word documents takes place in the user controlled environment of his or her local personal computer. To do this, the user uses the personal computer to access the MS Word content creation software application. This application automatically opens its default template when creating a new document. This is normally the “blank.dot” template, which does not contain visual content. However, it does include page layout information, template identification and text formatting styles.

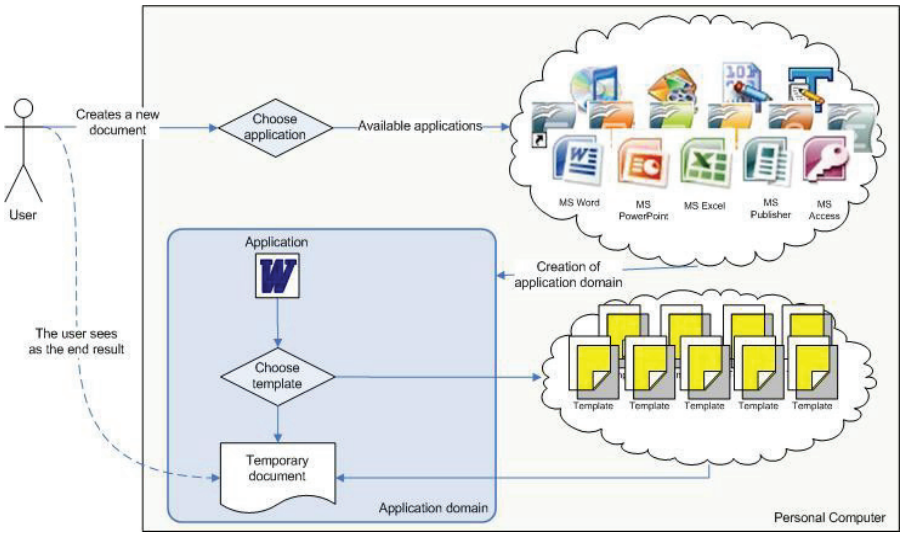


Figure 36: Creating a new stand-alone document

After the template is opened and presented to the user through the graphical user interface, the user is allowed to make changes to the new document. This is where the user first experiences creating document content. Here the user is allowed to use his or her creativity to develop the new document content and present its intellectual content.

Saving the stand-alone document

When the user gives the application the command to save, a number of actions are automatically performed:

- If there is no "Title" element recorded, then an algorithm is executed to generate this element. This algorithm collects data from the first line of text. The "Title" element is also used as the default document name. The file name may be changed, although the "Title" element is not automatically changed.
- The system clock is used to generate the "Creation date" element. If the user has printed the document, a "Last printed date" element is included with the collection of data from a temporary recording of the system clock at the time of printing.
- The application's user profile is used to populate the "Author" and "Company" elements.
- Technical metadata are generated by algorithms that analyze the document to retrieve entities for elements such as the number of "Characters," "Words" and "Pages." Other technical elements are collected from the template including page size (e.g. "Letter" or "A4"), margins and orientation ("Landscape" or "Portrait").
- All the metadata are placed within the document's metadata section.
- The intellectual content included by the template and the user (excluding metadata) is placed in the main document section of the document code. Extensive formatting descriptions are included so that all the properties of the document are kept. This includes text style formatting, language, imported content, etc.
- The document format extension is automatically changed from the template format (".dot") to document format (".doc").

This shows that there are a number of different factors that influence the content of each document's metadata elements and document content:

- The actions performed by the content creation software application
- The document template
- The user's performed actions
- The application's user profile
- The system clock
- The application's metadata generating algorithms

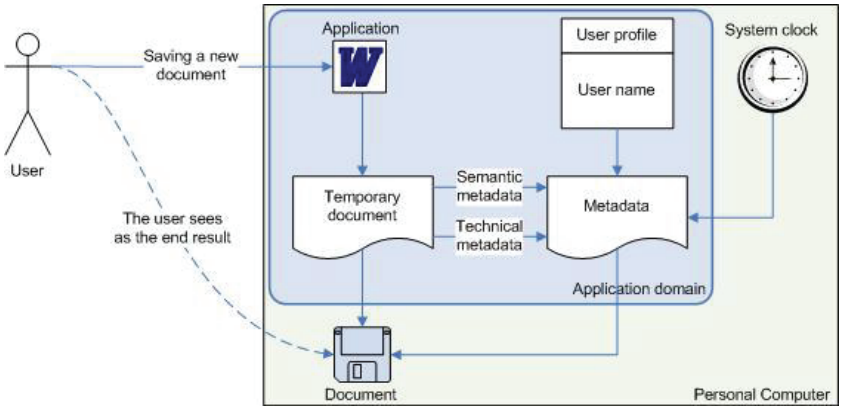


Figure 37: Saving a new document

Editing an existing stand-alone document

Based on the saved document, all the characteristics of the document should be retrievable from the document code. When opening an existing document for editing, the document code is used to bring all the document's characteristics back into the application's domain. The main document is presented to the user ready for editing. Selected metadata elements and their entities are presented though the graphical user interface, normally the pages element (e.g. "Page: 3 of 5"), Words (e.g. Word: 680) and Language (e.g. English (U.S.)). The entities for these elements are automatically updated as the user edits and navigates within the document.

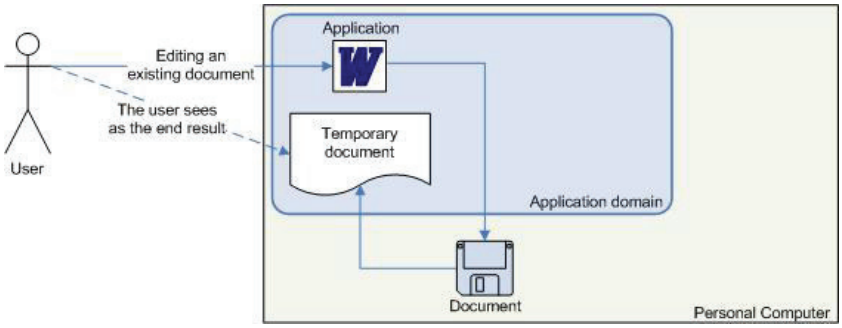


Figure 38: Editing an existing stand-alone document

Re-saving an existing stand-alone document

If the user gives the command to re-save the document, then a number of actions are performed to place the application's information about the document back into the document code. However, this saving process is not identical to the first time the document was saved:

- The title-generating algorithm is not executed since the document already has a metadata "Title" element. User-specified updates of the visual title of the document are not used to update the existing, embedded metadata "Title" element.
- The system clock is used to generate the "Modified date" element. If the user has printed the document since it was last saved, then the "Last printed date" element is updated.
- The application's user profile is used to update the "Last Author" element.
- The application once again executes an algorithm to collect and update the existing, embedded technical metadata elements.

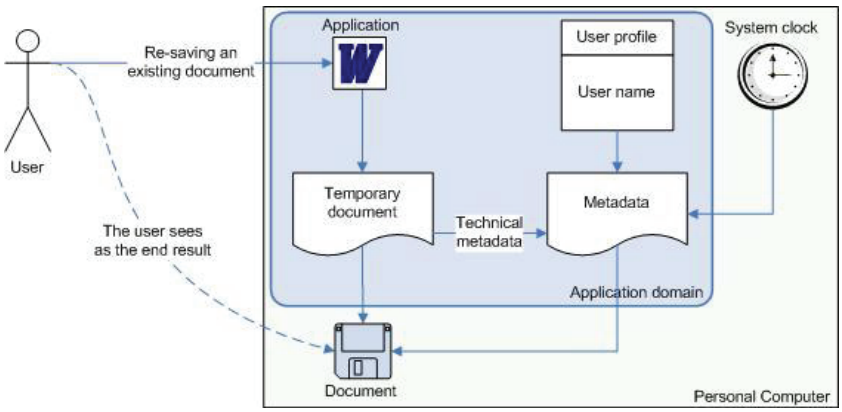


Figure 39: Re-saving an existing stand-alone document

Converting a stand-alone document

Many document creators choose not to publish their original documents. The reasons for this may include a desire to restrict usage and editing opportunities, and to ensure that the document is presented in a specific way. There are multiple ways in which a conversion can take place.

Within the case LMS, 87% of PDF documents were confirmed converted using a converter application running on the user's own computer, 7% used an online web application, 2% were scanned print-outs and 4% were missing "Producer" metadata. A total of 137 applications and application versions were recorded as having been used. Converting PDF documents using a web application differs from traditional applications

by requiring the user to store the original document before the conversion process can take place. Documents that are not stored before being converted (on the user's computer) do not go through the initial storage process, and hence the metadata-specific storage processes described in Chapter 4.1.3 are not necessarily executed. This increases the uncertainties regarding the content of the converted document's resulting metadata. The remainder of this chapter presents a conversion process as if executed from the user interface of the original document format's native application.

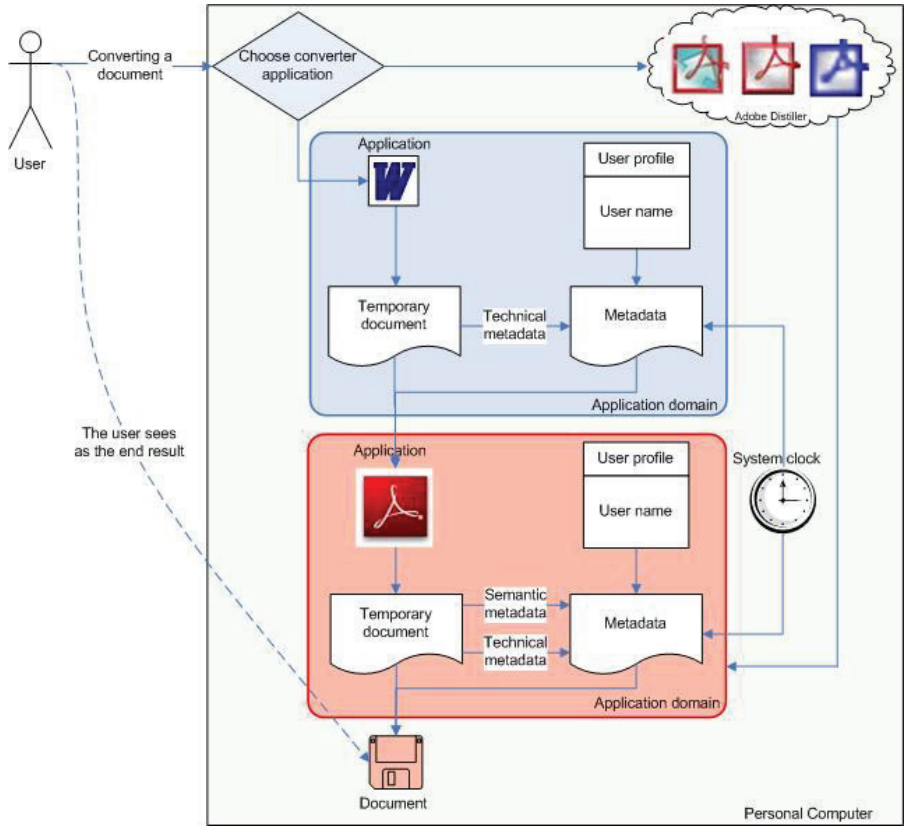


Figure 40: Converting a previously saved document

When the user gives the command to convert a document into a PDF document, this starts a new sequence of events. The document content and metadata are collected as if the document were to be saved (see Chapter 4.1.3) or re-saved (also see Chapter 4.1.3). However, instead of placing these data in a document, they are transferred to the domain of the converter application. It is then up to the converter application to decide what

should be kept as document content and metadata, and what should be changed. In this process the user may be allowed to make adjustments, e.g. specify security restrictions.

The main task of the conversion application is to generate a PDF document with a visual appearance as similar as possible to the original document. Since a conversion process changes the characteristics of the document, many of the embedded metadata elements do not reflect the converted document. It is therefore common practice for the embedded entities to be discarded and replaced by metadata generated by the converter application. As with the original document creator application, the converter application can collect data from a range of data sources:

- New semantic metadata are created based on another "Title" algorithm.
- New technical metadata are created based on the new technical characteristics of the document.
- The converter application's user profile is used for creating "Author" elements. Online converter services commonly use alternative data to be included in the "Author" element.
- Some converter applications allow the user to make corrections to the semantic metadata elements.
- The system clock is used to give a new time of when the converted document was created.
- The document content is re-formatted to the new document format. Existing non-visual formatting (e.g. formatting styles and language tags) is discarded.
- Finally the new document content and the new metadata are placed as document code within a new document.

Converted documents therefore reflect both the original document creator application and its application domain, and the application and the application domain of the converter application. As a result, there can be extensive differences between the content of the original document and the converted document. This reflects both the document's metadata and the content of the main document content section.

4.1.4 Summary

There is a clear distinction between documents created in the system controlled environment and stand-alone documents created without system enforced control. In the system controlled environment, the user is required to use system-specific applications that are not influenced by the user's personal computer or local software. All documents are based on predefined, system-specific templates. The application can enforce mandatory elements and restricted value spaces. To some extent, such applications can validate text-based entities provided by the user. Through log-in features, the system has full control over who the user is, the sections in which the user is allow to create documents, and hence the context in which new documents are created. This does not assure that all data sources from the system controlled environment are correct, high quality entities. However, the system controlled environment ensures consistency in the created documents while avoiding local interpretations and variations. This ensures that countermeasures can be effectively enforced if false content is detected.

Stand-alone documents can be created in an infinite number of ways. The source code of these documents reflects the computer system of the creator, the content creation software, the templates that were used and the actions performed by the user. Validation of the user's actions is not undertaken. Converting documents between non-compatible document formats further increases the uncertainties regarding the document code. As a result, stand-alone documents can be quite diverse with different document codes, even though the visual appearance of the documents is identical. In order to find common structures and consistency within the pre-study dataset of stand-alone documents, this research examines entities from such documents in Chapter 4.2.

4.2 Quantitative element analysis

This chapter analyses selected embedded metadata elements from published, stand-alone documents. It is based on the stand-alone documents discovered though the initial analysis presented in Chapter 4.1.1.

Chapter 4.2.1 presents the results of the pre-study dataset. It is mainly based on the element types developed by Dublin Core and IEEE LOM, although other elements are also described if they are present. The pre-study dataset is based on courses that this researcher had access to as a result of his own course of study, or courses that were made available by PhD colleagues. No documents created or published by this researcher have been included in the dataset.

Over time, this researcher was able to gain access to more course sections, spanning a range of the university's subject courses. After the pre-study, this research built a more extensive dataset that was the basis for the qualitative analysis presented in Chapter 4.3. The final dataset proved to contain properties that differed from the pre-study dataset. Chapter 4.2.2 presents the specific elements from the final dataset that differed from the pre-study results.

4.2.1 Uploaded stand-alone documents as part of system documents

The quantitative analysis was performed on documents downloaded from the case system. The documents' native content creation software can present metadata that are created when opening the document and not present in the document. This research therefore used a dedicated metadata harvester application to obtain embedded metadata from the documents without opening them.

The pre-study dataset

Stand-alone documents are not changed when uploaded to the case LMS. The documents therefore keep their initial properties, with the exception of the "Created date" element. When uploading stand-alone documents, the file name and file size are automatically harvested and displayed as part of the LMS document type. From the collection of 424 LMS documents, 289 stand-alone documents were retrievable. As these documents were gathered during the pre-study phase of this thesis, these documents are referred to as the "pre-study dataset". These documents were downloaded from the LMS and analysed.

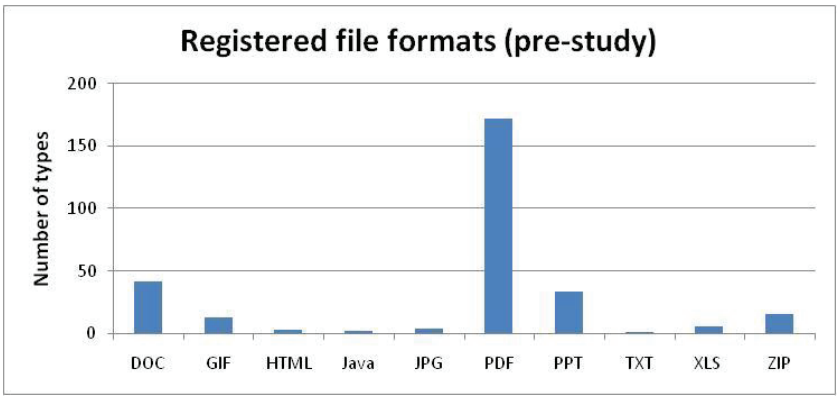


Figure 41: The pre-study stand-alone document format types (number of files for each document format)

This research has chosen to treat compressed documents as one compressed document, rather than as the number of uncompressed documents, because the document was shared as compressed. The statistics show that the majority of published stand-alone documents were of the PDF document format. Fully 59.5% of the published stand-alone documents were PDF documents, followed by Word documents (DOC and DOT) with 13.8% and PowerPoint (PPT and PPS) with 14.7%. DOC and DOT and PPT and PPS documents have been analysed as one document format since they are identical. The different document format names are used by applications to identify how the document is intended to be used when opened: DOT documents are templates that will be stored with the DOC file name after being edited. PPS documents should be opened as a slideshow in full-screen slideshow mode.

The stand-alone documents submitted are diverse, ranging from being based on predefined official administrative templates created by university employees, to documents without any apparent structure created by students on private computers. Examples of document appearances can be seen in Figure 35, Figure 51, Figure 57, Figure 62 and Figure 65. This dataset differ extensively from other AMG-related projects.

Some content creation software generates metadata elements without entities. Elements without entities (content) do not provide a value. Empty elements have not been collected or analysed. Some document formats section the metadata. Metadata elements located in a section are presented with their section name first, e.g. “EXIF. Date Time Original” from JPEG images and “DC. Title” from the Dublin Core (sub)-section of PDF documents. Elements that are not sections are referred to as “General elements.”

Educational metadata

No documents contained dedicated educational metadata. The metadata elements of the IEEE LOM schema’s “Educational” section would therefore have limited ability to

harvest entities from these documents. The exception is the element "Typical Learning Time," which to some degree can be regarded to be the same as the playing time of a movie or the length of a slide show if there is a timer for the slide show. No video document formats or slide shows with a timer were found in the pre-study dataset. This shows that there is a need for using alternative data sources for generating such elements, e.g. by using context information as described in Chapter 2.3.1.

Common metadata elements

Some stand-alone document metadata can be collected from the file system. Hence, these elements are generally present for all stand-alone documents regardless of their format or other metadata content:

Table 7: Common stand-alone document metadata elements

Element	Content	Example
Name	The file name	Husleier september 2006 til studenter.xls
Full name	The file name and its physical location as presented on the user's computer	e: \Husleier september 2006 til studenter.xls
Short name	The file name with a maximum of 8 characters and file format extension	HUSLEI~1.xls
Extension	The document format	XLS
Creation	When the document was created	2006/10/26 08:20:26
Last Saved	When the document was last saved	2006/10/26 08:20:26
Size	How many bytes the document consists of	16896

Some document formats contain the "Creation" and "Last Saved" elements as a part of their embedded metadata. If such metadata were present, then the harvester application automatically uses these data instead of the file system's data. This reflects the MS Office document formats. PDF documents use synonym element names: "Creation Date" and "Mod Date." PDF documents are therefore presented as containing all four date elements. Of these, the "Creation" and "Last Saved" elements only reflect the time at which the document was downloaded to this researcher's computer.

Semantic elements

Title element

All Word and PowerPoint documents contained a "Title" element. So did 83.7% of PDF documents and 33.3% of HTML documents. No other document formats were observed to contain a "Title" element. This includes Excel documents. Applications such as MS

Word[8], MS PowerPoint[9] and Adobe Distiller[10] automatically generate "Title" elements for created documents. There are therefore four potential creators of the "Title" element: the user, the template creator, the original document creator application and the document converter application. Selected PDF documents contained multiple metadata "Title" elements because they used a General element section and a RDF-based section containing DC, PDFX and XAP metadata elements.

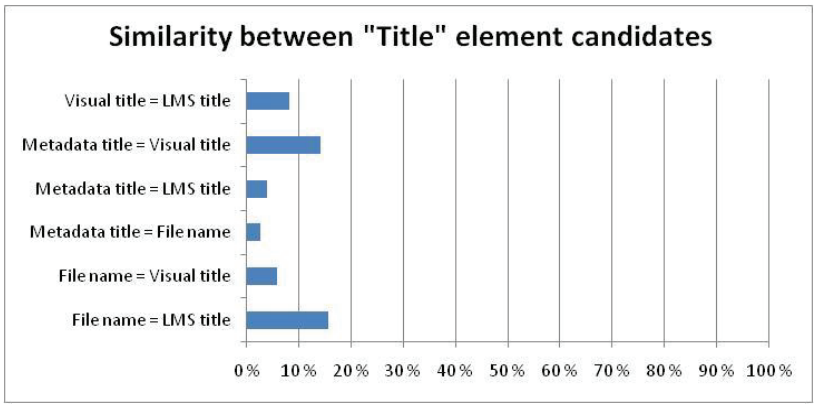


Figure 42: Similarity between "Title" element candidates (PDF, Word, PowerPoint and Excel document formats)[11]

The title presented when viewing the document through its standard user interface or print-out is referred to as the visual document title. There were extensive differences between the embedded "Title" entities and their visual titles. Common "Title" element content includes standard values, such as "Document1," and commercial content from online PDF converter applications. These elements do not reflect common metadata schema definitions nor are they representative of the visible content of the document.

[8] Application versions up until, but not including, MS Word 2007. MS Word applications were recorded as having been used to create 100% of the dataset's Word documents.

[9] Application versions up until, but not including, MS PowerPoint 2007. MS PowerPoint applications were recorded as having created 98.9% of the PowerPoint documents. Just 1.1% of the documents were recorded as without content creation software name metadata.

[10] Application versions up until Adobe Distiller 4.01. Adobe Distiller was recorded as creating 63% of the PDF documents.

[11] This comparison was only performed on file formats that can contain or retrieve all the candidate data sources. Hence JPG, Java, GIF and ZIP file types were not a part of this comparison.

The numbers displayed in Figure 42 indicate that the different candidate data sources contain very different data and that these data sources frequently differ from the visual document title. With a degree of similarity of only 14.1% between the visual and metadata titles, there is an extensive need to generate higher quality metadata "Title" entities. Efforts to do this are presented in Chapter 4.3.3.

A strong connection between candidate elements was discovered for GIF document images. Here 92.3% of the documents had the identical file name and LMS Title. This shows that there are differences between document formats regarding the correctness of using candidate data sources.

Creator element

Only the document formats for PDF, Word, PowerPoint and Excel contained elements that reflected the document creator. A range of elements was found that reflect this element, including "Author," "DC. Creator," "Last Author," "PDF. Author" and "XAP. Author." Validating the entities for these elements is a challenge, since only 45.9% of PDF, 22% of Word, 30.3% of PowerPoint and none of the Excel documents contained a visual creator name upon which a comparison could be based.

Twenty-seven PDF documents (15.7%) contained extensive amounts of false data in the "Author" element, e.g. *"Lars Edvardsen) /Creator (PowerPoint) /CreationDate (D:20060329110418+02'00'"* [12]. This text string presents the "Author" element at the beginning, before the "Creator" (application) element. Such false formatting is an issue that is found in all the PDF documents created using the application (metadata element "Producer") "Mac OS X."

All PowerPoint documents contained the "Author" and "Last Author" elements. These elements were the same 72.7% of the time. Only 15.2% of the time did one of these elements match the visual creator's name. Fully 18.2% of these elements were the same as the LMS publisher name. All Word documents contained the "Author" and "Last Author" elements. These elements were the same 73.2% of the time. However, because only 22% of Word documents contained a visual creator name element, comparisons are difficult. Only two documents had entities equal to their visual creator name and the LMS publisher name. All Excel documents contained the "Author" and "Last Author" elements. Of these, only two were the same. None of the "Author" or "Last Author" elements were the same as the LMS publisher name. Entities from Word, PowerPoint and Excel documents frequently contained name shortenings, software license registration user names, and default values.

For PDF documents, 76.2% had a "Creator" element, while 45.9% had a visual author. However, only 1.7% of these elements were the same. These extensive differences were influenced by entities that were altered by the converter application. The most commonly used conversion application (Adobe Distiller) has been observed by this researcher to discard embedded creator information, replacing it with its software

[12] The author names have been changed to make the real author(s) anonymous.

license registration user name. Such actions were also performed by online converter services where commercial content was included instead of the user's embedded metadata. All told, 8.1% of the visual author records were the same as the LMS publisher name.

There is much uncertainty regarding the Creator element for these document formats, partly because there are so few documents with a visual creator name, and elements based on user names and commercial content rather than user names. The dataset also contained many different ways of writing author names, such as with a surname, excluding middle names or without a first name. As a result of this, the Boolean comparisons undertaken in this chapter have not been able to distinguish correct and false elements. Determining this requires a deeper comparison between these elements, where manual judgment must be used in order to determine equality. This is has been done in Chapter 4.3.2.

Subject element

Three PDF documents contained a Subject element. These entities were all commercial content from an online PDF-converter service, and hence were all false.

Description element

No documents contained a "Description" element. The Word, PowerPoint, Excel and HTML documents contained the element "Comments," which can be used in the same way. No documents were found with entities for their "Comments" elements. This element seems not to be in use. Instead, the LMS is used to give document descriptions:

Table 8: Documents described in the LMS

	DOC	PPT	XLS	HTML	GIF	PDF	ZIP
Described individually	4.9%	15.2%	20.0%	33.3%	7.7%	7.6%	0%
Described as part of a LMS document	12.2%	0%	0%	33.3%	0%	26.2%	26.7%
Included in a blank LMS document	12.2%	0%	0%	0%	0%	2.3%	0%
Described either individually or as part of a LMS document	17.1%	15.2%	20.0%	66.6%	7.7%	33.7%	26.7%

Keywords element

"Keywords" can be included for Word, PowerPoint, Excel, HTML and PDF document formats. Only three PDF documents included "Keywords" elements with one or more entities. All of these were commercial content from an online converter application. All entities were hence false.

Publisher element

No documents contained a "Publisher" element. However, since all stand-alone documents were published though the LMS, publisher information can be collected from the LMS. The data source for the publisher element is regarded as a trusted data

source since the publisher needs to log in to the LMS before being allowed to publish a document.

Contributor

No documents were discovered with a "Contributor" element.

Technical elements

Date elements

An analysis of the created date element was only possible for stand-alone document formats that contained an embedded created date element. This was the case for MS Office and PDF documents. The other document formats proved to not contain a created date element(s). This was the case for TXT, HTML, GIF, JPEG and ZIP documents. These document formats were given elements with entities from the harvester application by collecting storage information from the file system. This assigned time reflects when the documents were downloaded to the computer used for this thesis, and not the actual creation date. The created date for these document formats has therefore been regarded as corrupted and has not been analysed. JPEG documents can contain date metadata from their EXIF metadata section. However, no JPEG documents contained this type of metadata section.

A range of date elements can be collected from PDF, Word, PowerPoint and Excel documents. This includes "Created," "Last Save," "Last Access," "Last Print," "Creation Date" and "Mod Date." From this list, the "Created" and "Creation Date" elements and the "Last Save" and "Mod Date" elements are synonyms. The "Last Access" element refers to the date when the document was last accessed. This date is therefore the same as the time at which the metadata were extracted. As such, this element does not provide value for this research. The "Last Print" element has not been analysed since it reflects usage information and is not document description metadata.

Less than a handful of documents contained a visible date element. The dataset was therefore regarded as too small to analyse.

Two formatting issues were discovered regarding the date entities from PDF documents:

- Two PDF documents (1.2%) contained "Creation Date" elements with entities that were falsely semantically formatted. The entities were dates, though not formatted like other PDF documents. These elements were collected from an old version of the PDF format (v1.2). No other PDF documents used this version of the document format.
- Twenty-seven PDF documents (15.7%) contained extensive amounts of false data in the "Author" element (see Chapter 4.2.1). These data included the "Creation Date" element.

Both these issues can be identified based on the document format version or the producer element (application version). It would be possible to adapt the AMG

harvesting algorithm to identify the specific application versions and document format versions in order to execute custom algorithms to perform corrections to the date entities. For the remainder of this chapter, this research has treated these elements as if they were correctly formatted.

Entity for the "Creation Date" element were missing for 8.7% of the PDF documents, while 77.7% of PDF documents contained "Creation Date" and "Mod Date" elements that were the same. None of PowerPoint documents and 22.0% of Word documents had the same elements. This show that these document formats are used differently: PowerPoint and Word documents are being worked with and re-saved multiple times before being published. In contrast, a large portion of the PDF documents are converted into this format after the editing process has ended. A large portion of the PDF, PowerPoint, Word and Excel documents contained entities that indicated that they were published the same day they were created or modified. This was true for 50.6% of PDF documents, 57.6% of PowerPoint documents, 80.5% of Word documents and 80.0% of Excel documents. According to the metadata, the oldest document in this dataset was from 1997.

All date entities created for stand-alone documents are based on the timer (clock) of the user's local computer. There is no information stored as part of the document or from the LMS that can confirm that this timer was correct when metadata were generated. The correctness of these entities cannot be confirmed. However, a few elements can be used to determine if entities are false. These actions can confirm if selected entities are false, though they cannot confirm if the entities are correct. This is true for comparisons between:

- Conflicting document entities: The "Created" and "Modified" elements. A document cannot be modified before it is created.
- Conflicting stand-alone document entities and LMS document entities: "Created" or "Modified" after the document was published to the LMS. The LMS does not allow stand-alone documents to be created or modified within the LMS. Hence, this situation cannot occur.

Fully 5.2% of the PDF documents had entities indicating that they were modified before being created, while 3.2% of the PDF documents had entities indicating that they were published before they were created or last saved. This situation was also found in one PowerPoint document. One Word document was recorded as modified after the document was published. These observations confirm that date entities from stand-alone documents cannot be fully trusted as quality metadata.

Format

The document format can be identified by the file name extension. This data type is available from all stand-alone documents as part of the descriptive elements that can be collected from the file system of the computer system in which the document is stored.

Type

No documents were discovered that contained this element, although it is possible to infer the "Type" element from the "Format" element. This is because most document formats have a dedicated primary usage area. These usage areas can be used to give default entities based on the value space of the "Type" element from the Dublin Core schema [29], e.g.:

- Text: DOC, TXT, PDF
- Dataset: XLS
- Moving image: Animated GIF
- Still image: JPEG, GIF
- Interactive Document: HTML, PPT
- Software: Java
- Collection: ZIP

Identifier

Close to half of the PDF documents contained internal identifiers. These identifiers described the document as an identified object and for sub-content (such as images) that were found in the PDF document. These metadata can be collected from the "RDF. About" and the "XAP. MMDocumentID" elements. These elements contained entities based on the Universally Unique Identifier (UUID) standard [131], such as "uuid:ab14519a-2206-4e38-847f-5742eb64aa7d." This standard was designed to allow users to create documents on their local computers with a globally unique identifier without central coordination. The Word, PowerPoint, Excel and JPEG document formats also support use of this or closely related identifier schemas, though no such content was discovered in the pre-study dataset.

Language

No documents were discovered containing embedded metadata relating to the language of the intellectual content of the document.

Source

No documents were discovered containing embedded metadata relating to the "Source" element as defined by the Dublin Core schema or "Relation" of the IEEE LOM schema.

Relation

No documents were discovered containing embedded metadata relating to relationships with other documents, aside from the HTML references to format and schema definitions.

Coverage

No documents were discovered containing embedded metadata relating to the "Coverage" element as defined by the Dublin Core and IEEE LOM schemas.

Rights

User rights and security restrictions can be specified as part of the metadata of PDF documents. All the registered documents contained rights metadata indicating "no restrictions".

Characters, Words, Pages and Slides

Measuring the amount of intellectual content in a document is close to impossible, because there are so many different ways to express yourself, and there are an equal number of different ways in which the document itself can be understood by the user [90]. The numbers of characters, words, pages or slides are among the few technical elements that can be visually verified by comparing the document's metadata and their visual characteristics. These elements can offer indications regarding the quantity of the document's intellectual content and what type of document it is, as in a flyer, a brief paper, a term paper, a book chapter or a completed book.

The Word and PDF document formats contained metadata describing their number of pages. PowerPoint documents also included metadata regarding the number of slides and Excel documents regarding the number of sheets. There was agreement between the visual number of PowerPoint slides and Excel sheets and their entities.

Table 9: Elements available in the different document formats

Element	PDF	Word	PowerPoint
Characters		X	
Words		X	X
Pages	X	X	X
Slides			X

The Word documents showed an error rate of 69% for the software's embedded "Pages" element. All these issues resulted from too few pages being recorded in the metadata. Most Word documents had entities indicating a single page, with 17 of the 20 documents with the highest number of characters and words all recorded as having one document page. This indicates inconsistency within the metadata. All the Word documents with page errors were created with MS Word 10 or 11 (otherwise known as MS Word 2002 and MS Office Word 2003). These applications stood for the majority of Word documents with a correct number of pages as well. These applications can therefore create both correct and faulty metadata.

The error rate for the embedded "Pages" element of PDF documents was 25%. Unlike the Word documents, the PDF metadata had entities with numbers that were too high and too low: 7% were too high, while 18% were too low. There were extensive differences between the creator applications for these documents. The most commonly used application (Adobe Distiller) had an error rate of 3%, while documents created by Mac OS X had an error rate of 45%.

The PDF documents also showed the additional challenge of having multiple visually present slides, or "logical pages" per "technical" page. The amount of information per slide does not change when multiple slides are printed on one page. With up to 9 logical pages or slides per technical page, there can be a substantial difference between what is

perceived by the users and the number of pages that this element indicates. The number of pages element can therefore be misleading even though it is technically correct. Making a distinction between the number of visual slides or pages and technical pages by using "qualifiers" (as in Dublin Core) or separate elements would avoid this issue. The inclusion of problems caused multiple pages increases the error rate for PDF documents to 32%.

This research conducted an in-depth analysis of the issues around potentially false "Characters," "Words," "Pages" and "Slides" elements, found in Chapter 4.3.1. This required an in-depth analysis of stand-alone documents. Since the number of characters, words, pages and slides are closely logically related and visually verifiable, these elements were analysed together.

Template information

The Word and PowerPoint document formats can contain metadata that presents the identification of the template upon which it was based. For example, 95.2% of Word documents were based on the blank template "normal.dot." This is the template that is the default for MS Word when a new document is created; see Figure 32 (p. 94). This template does not include any visual content, which indicates that users commonly use the blank template document and adapt it to their specific needs instead of using a task-specific template. This has an additional effect on the document code in that template sections are not formatted with template-based styles. As a result, there are numerous different usages and little consistency within the dataset.

The blank, default template for PowerPoint documents ("normal.pot") contains visual template sections as presented in Figure 33 (p. 94). A direct consequence of using this template is that more users take advantage of available template sections, as has been documented in Chapter 4.3.3. The name of this default template is not stored as part of the document code. Instead, only alternative templates are recorded, if in fact they are used. A template was recorded for 18.2% of PowerPoint documents, but all were different from the default "normal.pot" template.

The official NTNU document templates were published only in a very limited way. The dataset contained only a single Word document and two PowerPoint documents that were based on a NTNU template.

4.2.2 The final dataset

In total, this research analysed the content of 166 course sections in the case LMS. This counts for approximately 11% of all courses at NTNU [88]. In total 3483 stand-alone documents published from these courses. These documents are referred to as "the final dataset".

The pre-study showed that much content is reused when a course is offered multiple times, e.g. in the spring of 2005 and later in the spring of 2006. To avoid duplicate documents created by the same publishers, this research excluded courses with identical course names, with only the most recently offered course analysed. In addition, courses that were related to this research were excluded from the final dataset. In total 32

courses were excluded from analysis due to these two issues. This includes some courses that were used in the pre-study phase.

General statistics

The final dataset consisted of 3483 documents. There were a total of 41 different stand-alone document formats that had been published. Of these, three document formats dominated the statistics: Adobe PDF documents (1943 documents, 55.8%), MS Word (DOC) (745 documents, 21.4%) and MS PowerPoint (PPT and PPS) (475 documents, 13.6%). These document formats comprised 91% of the documents in the dataset. This research effort was thus concentrated on these document formats.

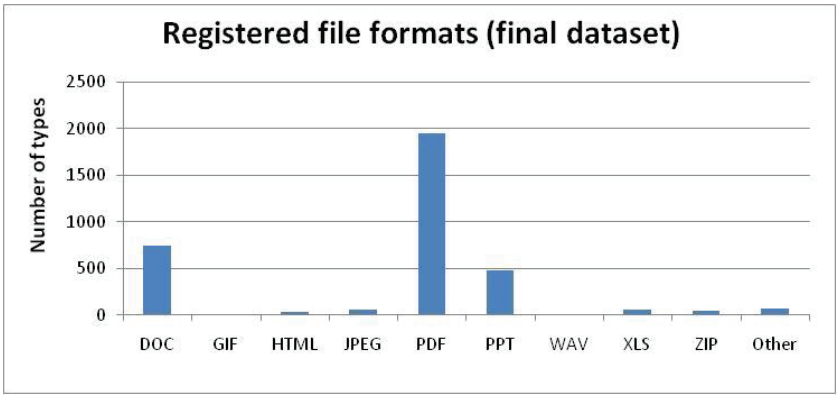


Figure 43: Stand-alone document format types (number of documents for each document format)

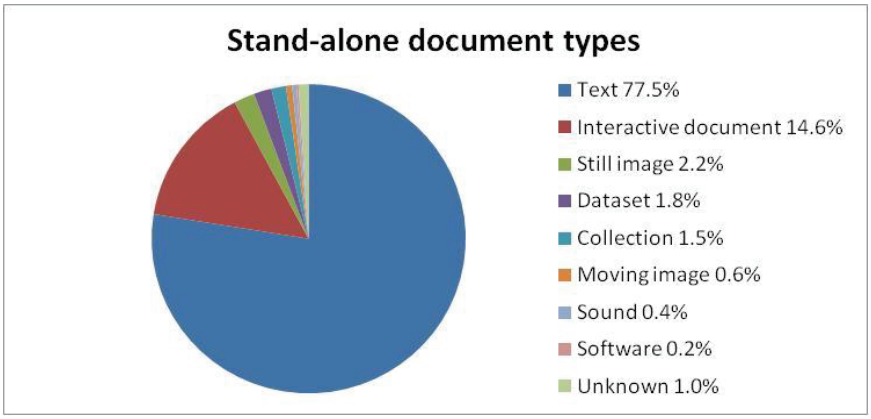


Figure 44: Stand-alone document types based on the document formats' primary usage area and Dublin Core "Types"

Figure 44 shows the published document types based on the types as presented in Chapter 4.2.1.

Videos are included in the LMS only to a very limited extent. Instead there is extensive use of hyperlinks to external video sources. To create such references, the LMS "Link" document type is frequently used.

In total, 164 different element types with entities were harvested from the documents in the final dataset. These elements were located in the sections presented in Table 10. A number of these elements reflected the same issue of interest. For example, at least 5 elements reflected the "Title" element[13]. Even when duplicate elements reflected the same document, all entities do not have to be identical. This issue is further discussed in Chapter 4.3.

Table 10: Recorded elements

Element section	Number of elements
General elements	33 elements
Dublin Core	5 elements
EXIF	40 elements
IPTC	12 elements
PDF	11 elements
PDFX	15 elements
Photoshop	4 elements
RDF	8 elements
TIFF	11 elements
XAP	21 elements

All stand-alone documents were given at least seven elements regardless of document content, as discussed in Chapter 4.2.1. The average document contains 21.35 elements, with as many as 61 elements collected as the maximum. The majority of documents contained between 16 and 30 elements, see Figure 45. The use of elements varied extensively between document formats. The PDF, Word, JPEG and PowerPoint document formats contained the greatest number of elements, see Table 11.

[13] "DC. Title," "iptcbylinetitle," "PDF. Title," "Title" and "XAP. Title" plus possibly "Name."

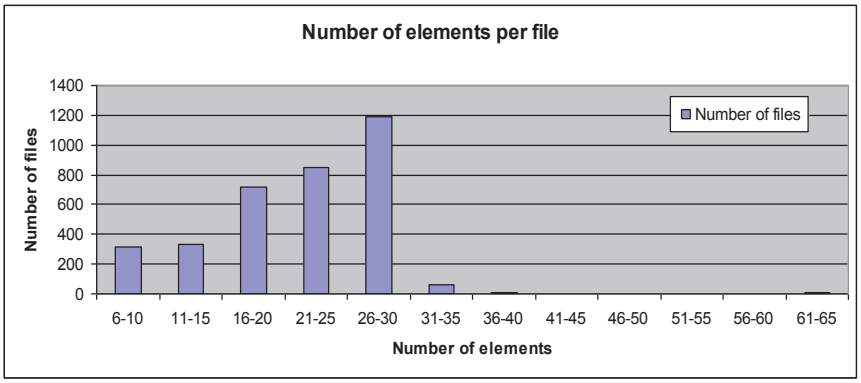


Figure 45: Number of metadata elements collected per stand-alone document

Table 11: Number of elements per stand-alone document format

	PDF	Word	PowerPoint	Excel	JPEG	TIFF	GIF	TXT	HTML	All formats
Minimum	10	18	11	7	8	8	8	8	8	7
Mode	26	21	19	12	8	8	8	8	8	26
Maximum	60	24	24	15	61	8	8	8	10	61
Average	23.6	21.3	19.1	11.8	20.5	8.0	8.0	8.0	8.2	21.4
Median	26	21	19	12	8	8	8	8	8	21

In addition to the elements presented in Table 10, the most common elements were "Author" (76.8%), "Pages" (75.9%) and "Title" (73.8%). These are all elements that are common in multiple document format schemas. A number of the sub-schema sections presented in Table 10 refer exclusively to technical issues. For example, the EXIF, IPTC, Photoshop and TIFF sections only contained content referring to photo technical properties. The majority of sub-schemas were located in JPEG images and PDF documents. The TIFF documents present the same opportunities for metadata descriptions as JPEG documents. Still, TIFF documents only contained just above the minimum of metadata elements. No TIFF images contained TIFF metadata (!). Only specific PDF documents contained TIFF metadata (PDF documents can contain full-word TIFF and JPEG images). The entities included issues such as the camera brand, shutter speed, white balance and colour settings. These elements contain entities that this research cannot verify. These elements have not been included in the analysis efforts.

Use of elements with differ from the pre-study dataset

The "Description" element was not found in the pre-study dataset, although this element was used in the final dataset. The first chapter presents these observations. The "Keywords" element was observed in PDF documents that presented commercial content. The second describes other observations regarding this element based on the final dataset. More document formats were observed using identifiers. These observations are presented in the third chapter. The other elements analysed in the pre-study were in line with the final dataset. These observations are not presented, as they appear to be almost duplicates of the pre-study results.

The Description element

The final dataset contained a number of elements that reflected the "Description" element in the IEEE LOM schema and the "Subject" element in Dublin Core:[14].

One Word document (0.1%) contained a "Comments" element, which was a date, although no other information was provided with it. This limits the usability of this element since there is insufficient information to interpret the data. This date was not identical to any of the other embedded data elements. The document was based on an official NTNU template that does not contain this entity. This indicates that the user has specified this entity, though it is not possible for this research to determine what this entity refers to.

Nineteen PowerPoint documents (4.0%) contained a "Comments" element. These all referred to the document templates upon which the documents were based.

Twenty-four PDF documents (1.2%) contained a "DC. Description" element:

- Five entities were valid entities created by the user. These entities contained keywords from the subject at hand.
- Fifteen documents contained entities that were number codes (e.g. 725-403) or default values (e.g. WithoutName-7). These documents were created using the "Adobe PageMaker 7.0" application. The number code entities were all identical to the "Title," "PDF. Title," "Subject" and "DC. Subject" elements. No templates were recorded for these documents. However, based on extensive visual similarities, it appears that these documents were based on the same template, which was a building legislation template. There was only one section that was visually the same as the "DC. Description" entity. This section only contained strictly standardized number codes. The variations discovered in the "DC. Description" element was not found in this section. This researcher concludes that the user has specified this element, although it is not possible to conclude which element was the original or correct element. All the documents

[14] Elements: "Comments," "Notes," "PDFX Comments," "DC Description," "XAP Description," "Subject," "DC Subject," "PDF Subject," "PDFX EmailSubject," "Category," "iptccaption" and "iptcbyline."

were renamed to receive standardized document names based on the number codes (e.g. "725403").
- Three documents contained the entity "Image," which was automatically recorded by a scanner application.
- One document contained an entity with content intended for other elements[15]. This has been recognized as a problem for PDF documents that have been created using the PDF converter application included in the Mac OS X operating system. This results when the converter application specifies metadata that are not in accordance with the PDF standard.

Eighteen PDF documents (0.9%) contained a "PDF. Subject" element.

- Fifteen documents that were created using "Adobe PageMaker 7.0" contained entities identical to their "DC. Description" element.
- The three remaining documents were created using the most common PDF creator application "Acrobat Distiller 5.0 (Windows)." These "PDF. Subject" entities contained keywords derived from the subject at hand. One of these documents was created using a non-standardized driver. This was the only document that contained a "PDF. Subject" element, but no "DC. Subject" element.

A single PDF document (0.1%) contained a "PDFX. Comments" element. This was an extensive description of the actions performed by the user. This element was not repeated in any other elements, not even the "PDF. Comments" element. A commonly used application and application driver were used for document creation in this circumstance.

Keywords

Thirteen Word documents (1.7%) contained a "Keyword" element. All these elements referred to the document template that was used.

Seven PDF documents (0.4%) contained a "Keyword" element. All these elements referred to the document template that was used or commercial content from the converter application.

Two PDF documents (0.1%) contained a "PDF. Keywords" element. These elements referred to the document template used, and were identical to the "Keywords" element.

These observations confirm that the embedded "Keywords" element is not used by users. This element is instead used to distribute template information and commercial content. As the entities did not reflect the documents at hand in accordance with common metadata schemas, the embedded entities related to "Keywords" elements are hence of very low semantic quality.

[15] "Capturefile: C:\Documents and Settings\Administrator\Desktop\New England\1D\38AB1307.TIF, CaptureSN: 0000138A.014829"

Identifier

In the pre-study dataset, identifiers were only located in PDF documents. In the final dataset identifiers were located in selected JPEG and PSD image documents as well, as shown in Table 12. The percentage of use among PDF documents is almost the same in the pre-study and final datasets.

Table 12: Identifiers within stand-alone documents (both datasets)

	rdfabout	xapMMDocumentID
JPEG	1.6%	35.5%
PDF	33.3%	53.6%
PSD	0.0%	100.0%

4.2.3 Quantitative Summary

In this chapter this research has presented an overview of what is commonly present in document files from NTNU's intranet. There are primarily Adobe PDF, MS PowerPoint and MS Word documents that are shared. Virtually no documents contained either an educational metadata description or an informative description. It is evident that the document authors and publishers use minimal efforts in giving the document a semantic metadata description. So few in fact have given a semantic description besides the Title element, that it is highly questionable if the documents authors and publishers are aware that such content can be stored as part of the document.

It is evident that a number of entities stored as part of the document is not created by the user. Technical elements including file format, a number of time and dates and the number of pages are typically automatically generated. We see a worrying issue here, as a number of the elements with entities created probably without the user's awareness, contains a number of flaws. We commonly find Title elements with little or no resemblance to the author created visible title. And we found Creator elements certainly not created by the "Creator". In some documents we found a number of several contradictive entities describing the same document.

There seems to be consistency in terms of what documents types that are shared, and to some extent how documents are created. However, there is considerable uncertainty regarding the quality of the gathered document metadata and regarding the awareness to metadata by document authors and publishers.

4.3 Qualitative element analysis

This chapter presents an in-depth analysis of problematic elements resulting from stand-alone documents, as described in Chapter 4.2. For this analysis a new dataset was collected, consisting of 3483 stand-alone documents from 166 different courses, referred to as "the final dataset". A random selection of documents was selected from this dataset for in-depth analysis in each of the following chapters.

Common content creation software generates extensive metadata descriptions of stand-alone documents. Chapter 4.2 described how several elements could not be verified or that there were uncertainties regarding synonym elements. Chapter 4.3 goes into more detail about specific elements that can be verified in order to determine the best candidate elements and data sources for generating desired entities. This chapter uses the final dataset and focuses on PDF, Word and PowerPoint documents, which make up 91% of the dataset. Chapter 4.3.1 presents an analysis of three automatically generated technical elements, "Characters," "Words," "Pages" and "Slides." Chapter 4.3.2 presents an analysis of the semantic element "Creator" (user), with multiple potential metadata creators. Chapter 4.3.3 presents an analysis of the "Title" elements from Word and PowerPoint documents. The chapter continues by presenting an alternative algorithm to generate "Title" elements, plus the result of using this algorithm. Chapter 4.3.4 describes how the language of the documents' intellectual content can be automatically determined without the need for evaluation of the document content.

4.3.1 The "Characters," "Words," "Pages" and "Slides" elements

The document content

This chapter presents an analysis of the "Characters," "Words," "Pages" and "Slides" elements, which are among the few automatically generated technical elements that can be visually verified for correctness. The analysis is intended to determine whether commonly used document creation applications generate high quality technical metadata. For this in-depth analysis, 245 PDF, Word and PowerPoint documents were selected at random. These elements are of special interest to this research because their entities are *always* automatically generated by the creator application, which means these elements can be used to visually validate the correctness of fully automatically generated metadata entities.

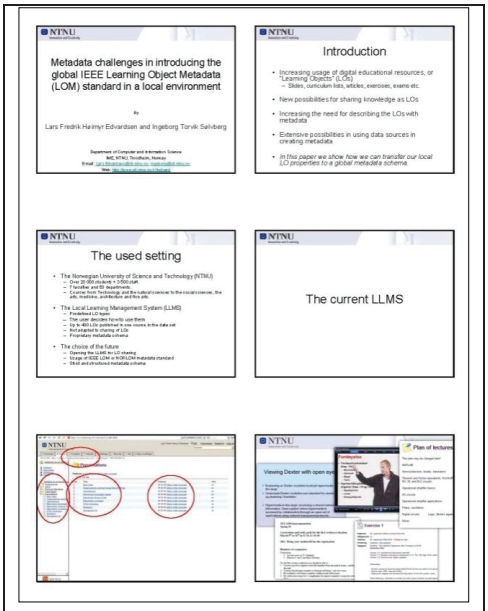


Figure 46: Example of a multi-page document with 6 logical pages on one technical page

A complicating factor regarding the number of pages is multi-page documents: Some document formats allow multiple logical pages on each printed page. A common example is when slide show presentations are printed with multiple slides placed on a single printed page. This reduces the number of pages in printouts, but does not reduce the amount of content in the document. The amount of information per slide does not change when multiple slides or pages are printed as one page. This

can cause a mismatch between the user's understanding of how many pages in the document and the amount of print-out pages that the document actually has.

Table 13: "Page" and "Slides" elements

	PDF	Word	PowerPoint
Technical page/slide errors	5%	66%	0%
Multi-page documents	21%	0%	0%
Document attachment	Yes	No	No
Security restrictions	Yes	No	No

Figure 46 shows an example of a document where 6 logical pages were placed on a single print-out page. Any reference to the amount of printout pages in the "Pages" element can therefore be misleading. This enables two types of "Pages" element errors, technical and logical errors.

- Technical errors occur if there is disagreement between the metadata "Pages" element and the visual number of pages from a printout.
- Logical errors occur if there is a mismatch between the correctly listed technical number of pages and the number of logical pages.

The analysis

This research examined a total of 41 different document formats collected from the case LMS. Of these, only Adobe PDF, MS Word and MS PowerPoint documents proved to have metadata schemas with embedded elements that related to the "Character," "Words," "Pages" and "Slides" elements, as shown in Table 9 (p. 110). Some PowerPoint documents contained the "Pages" element. This was unexpected since PowerPoint works with slides, not pages. Both these elements have been analysed. None of these document formats proved to contain metadata schemas that differed between the number of technical and logical pages. Only the technical number of pages has been included in these documents and their metadata schemas.

A total of 90.8% of the stand-alone documents uploaded to the LMS are in PDF, Word or PowerPoint document formats. PDF documents were the most common document format with 1943 documents (55.8% of the final dataset), followed by MS Word (DOC and DOT) (745 documents, 21.4%) and MS PowerPoint (PPT and PPS) (475 documents, 13.6%). Initially, 100 documents were randomly selected for analysis, resulting in 66 PDF documents, 22 Word documents and 14 PowerPoint documents. But the results of the "Pages" element analysis were so dramatic that an extended dataset was needed to validate the results. In total 243 documents were therefore analysed, of which 66 were PDF documents, 122 were Word documents and 57 were PowerPoint documents.

This research used the latest version of the document formats' native application to retrieve the documents' visual characteristics. These applications were also used to collect the entities that were presented through the "Properties" user interfaces of the

application. A dedicated metadata harvester application was used as the primary tool to collect embedded metadata, called "Metadata Miner Catalogue 4.2.20" [126]. A dedicated document counter application was used to extract the number of characters with and without spaces and the number of words, called "Any Count 6.0" [5]. Other applications were used to verify the correctness of the application results. Additionally, this researcher manually counted all characters and words.

Special characters and symbols were not included when the number of characters was counted. In order to avoid words like "A" and "B," a word was defined as consisting of two or more characters. This avoids having a number of meaningless "words" included in the counting. A consequence of this is that "I" and "å" ("to" in Norwegian) were excluded as words. Special characters, symbols and single letters were not counted as words. Consequently, the following data sources were used for each stand-alone document:

- The embedded "Character," "Words," "Pages" and "Sheets" metadata elements.
- The extractable data sources: The number of characters with and without spaces and the number of words counted by the counter application.
- The visual number of technical pages or slides.
- The visual number of logical pages per technical page.
- The document format's native application- the number of characters, words, pages and slides presented.
- The manually counted number of characters, words, pages and slides.

In addition, data were collected regarding the formatting of each document format in order to determine the possibilities and alternative data sources.

The "Characters" element

Only the Word document format contained metadata regarding the number of characters, although it is not clear from the format whether or not its definition of characters includes spaces.

Figure 47 uses the manually counted entities as a baseline (with the value "100 %") for comparisons against the other data sources. These statistics show that the data sources varied from counting only 44% of the correct number of characters, to 35% more or even 48% more when spaces were included in the count.

The entities gained from extracting the number of characters including spaces proved to be on average 17% higher than the number obtained from manual counting. The average number of embedded entities was also higher than the manually counted entities, which leads to the conclusion that the Word schema "Characters" element consists of the number of characters without spaces.

The entities presented through the application interface are not equal to the elements presented in the metadata document (!). This was confirmed by the use of test documents created for this research. The average results presented by the application and the extracted entities were on average slightly lower than results from the manual

count. However, as Figure 47 illustrates, a selection of entities was significantly lower than the manual count. An analysis of the documents showed that the entities that were harvested, the entities presented by the application and the entities that were extracted did not include text as part of:

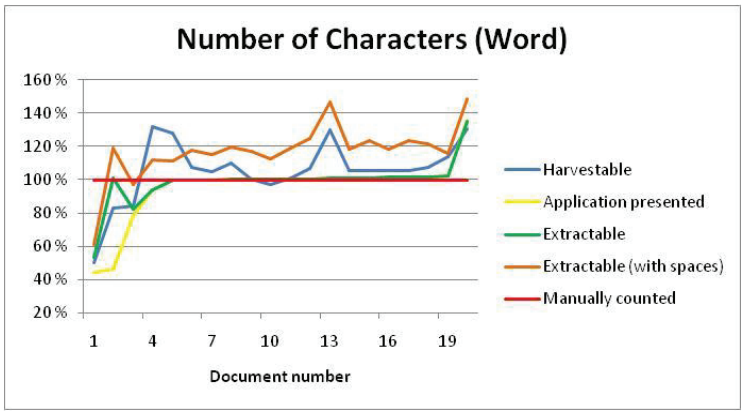


Figure 47: Number of Characters (Word documents)

Table 14: Issues affecting counting algorithms

	Issue
A)	Footnotes
B)	Endnotes
C)	Header
D)	Footer
E)	All other imported content

Documents containing content presented in Table 14 were given entities that were too low.

In addition, this research has observed three different approaches to what should be regarded as a character: The application-presented entities and the entities that were embedded included all text as characters. This resulted in a higher number of characters than what was actually correct. The extractor application generated entities based on the number of letters excluding special characters and symbols, and hence provided a better basis for determining the number of characters in the document. On average, the harvested entities were closest to the manually counted entities, although there were entities that were also too high and too low.

The "Words" element

There were extensive differences regarding how the applications and their document formats performed in regards to "Words" elements. These document formats have therefore been analysed separately.

Word documents

The different data sources provided a variety of entities for the same documents. None of the data sources were particularly accurate when compared to the manual counting efforts. This was the result of the same issues as with the "Characters" elements, presented in Table 14. The extracted entities showed to be the most similar to manually collected entities.

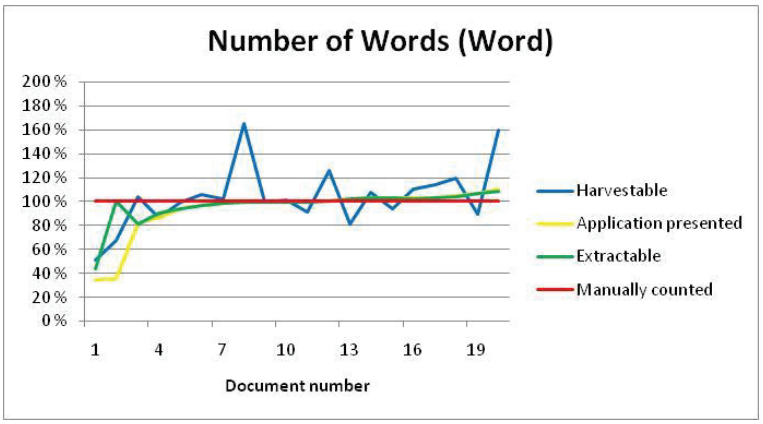


Figure 48: Number of words (Word documents)

PowerPoint documents

The PowerPoint documents were also strongly influenced by the issues presented in Table 14, but two additional issues caused variations in the different entities:

- PowerPoint documents contain a "Slide master," with template content that is presented on all slides and that is used instead of a header and footer in Word documents. The slide master appears to be used frequently, which resulted in fewer words being counted than what is visible.
- Not all applications included imported content when counting was undertaken. Only plain text content was counted. All other content were not counted.

Due to these issues, the application-based counting efforts did not tally enough words, on average.

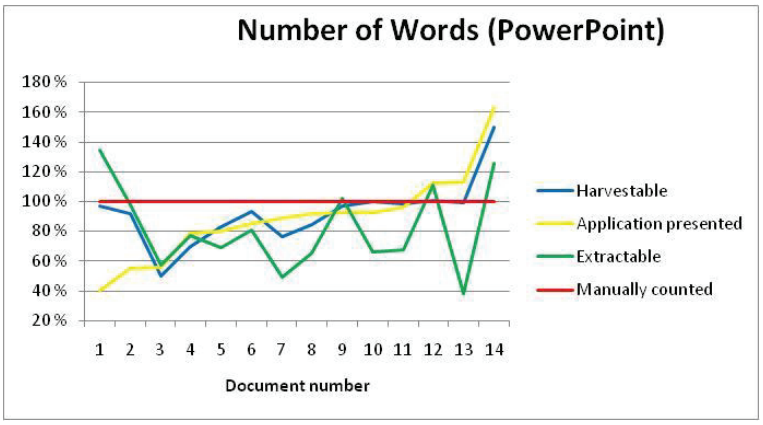


Figure 49: Number of words (PowerPoint documents)

The effect of this issue was especially visible for single slides with a great deal of imported content. Figure 50 shows an example of this, where much of the visual content is imported content:

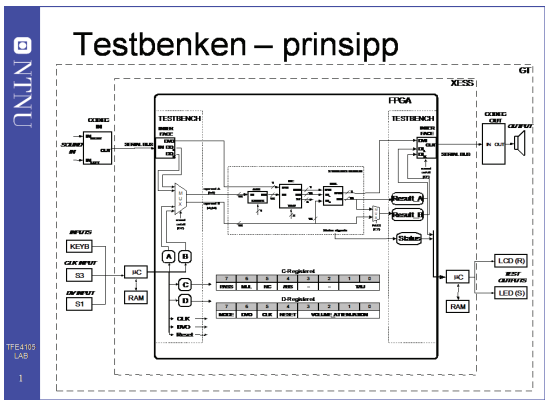


Figure 50: Example of PowerPoint document

Table 15: Entities collected from Figure 50

Data source	Entity
Harvested	0
Extracted	4
Application- presented	4
Manually counted	147

The heading (Textbenken – prinsipp"), course code (TFE4105 LAB) and slide number ("1") are based on a NTNU PowerPoint template, found in the slide master. The "NTNU" logo is an image and therefore should not be counted. The illustration was recorded as having been created as a Word document, though it is fully editable by PowerPoint: All the text-based content is editable. The manual counting efforts indicated that there were 147 words on this slide. However, the harvester did not locate any words, while the extracted and application-presented counts agreed that the slide

contained 4 words. These applications perceived the phrase “Testbenken – prinsipp” as if it were three words. All the applications used added an extra word to the “Words” entity. This occurred with *all* documents, even blank documents without *any* content.

Figure 51: First slide of a document with extreme results

Table 16: Entities collected from Figure 51

Data source	Entity
Harvested	14
Extracted	113
Application-presented	14
Manually counted	113

Inconsistencies regarding the counting efforts of the MS PowerPoint application were also observed with plain text content. In the slide presented in Figure 51, all the visual text is plain text, not imported content or images. Here the extracted entity matched the manually counted entity, but the harvested and application-presented entities tallied only 12% of the manually counted entity. Here the main text section was evaluated as containing 10 words instead of the correct 105 words. The string of numbers on the bottom of the slide was counted as one word, all together. The heading was correctly counted as containing two words (“Gruppe” and “5”). This research has not uncovered any documentation that describes why these sections have been counted in different ways.

Alternative methods of extracting the “Characters” and “Words” elements

This research involved experiments to determine if it was possible to extract more accurate entities. A simple application, called “PDF Reverser v01.01” [68], was tested. The application copies all text located in PDF documents into a plain text document. It confirmed that the text-based content of both Word and PowerPoint documents were accessible as plain text even though the document content needed to be extracted from a PDF version of the documents. Using the dedicated counter application on the plain text document confirmed that correct entities can be generated: In the case of the single slide from Figure 50, the PDF extractor returned 521 characters and 152 words. By filtering out the one-letter “words,” along with special characters and symbols, the results included 504 characters and 140 words. Manual counting resulted in 147 words. The

filtering did remove some two-character words that had been incorrectly split apart by the extractor application. Compared to the original application-presented entity and the original extracted entity, the correctness rate still increased from 3% (!) to 95%. By using an algorithm that performs a more accurate text extraction, this correctness rate can be increased.

The "Pages" and "Slides" elements

There were different issues that affected the "Pages" and "Slides" entities of PDF, Word and PowerPoint documents. Each document format has therefore been given its own subchapter: The first chapter presents PDF documents, the second chapter presents Word documents and the third chapter presents PowerPoint documents.

PDF documents

The analysis of issues regarding the "Pages" element for PDF documents is split into two sections, with technical errors in the first subchapter and logical errors in the second chapter.

Technical Errors

There were technical errors in 4.5% of the PDF documents (3 documents). These technical errors were caused by security-restricted documents, in which the "Encrypted" element was positive. These are documents where the user has explicitly specified that access to the content should be restricted. As a direct result of the restrictions the harvesting application has not gained access to all of the document's content data. These security-restricted documents also restricted access to a number of other metadata elements:

- **User information:** "Author"
- **Title:** "Title," "DC. Title"
- **Dates:** "Creation date," "Mod date," "XAP. Create Date," "XAP. Metadata Date," "XAP. Modify Date"
- **Application:** "Creator," "PDF. Producer," "XAP. Creator Tool"
- **Other element:** "DC. Format"

As a result of the security restrictions there are less embedded metadata available. Depending on the degree of security restrictions, there can be enforced restrictions on the ability to extract metadata as well, particularly if copying content in general is not allowed from the visually presented document.

The PDF document format allows multiple additional documents to be attachment to a single "master" PDF document. For example, the Adobe PDF Reference and Related Documentation consist of a single page PDF document with four sub-documents [4]. These documents consist of 1.334 pages. The master document only presents metadata about itself, which does not include its attachments. None of the PDF harvester or extraction applications used and tested by this research was able to identify these attachments. There can therefore be more content in a PDF document than what is presented though the embedded and extractable metadata. The document presentation application "Adobe Acrobat v8.1" presents a dialog box when PDF documents with

attachments are opened. By opening all the PDF documents in the dataset, this research has documented that none of these contained attachments.

All PDF documents without security restrictions had the correct number of technical pages. Still, the "Pages" element of PDF documents should only be regarded as partly reliable. This is a result of potential attachments that can be allowed in PDF documents. The entity for single PDF documents should be regarded as reliable. For security-restricted documents, no entities could be harvested. The number of pages can be extracted by parsing the document and then counting the number of visible pages. It is therefore possible to obtain reliable entities for security-restricted documents as well.

Logical Errors

An analysis of logical errors was undertaken after the security-restricted documents were excluded. These documents lack embedded entities against which to base a comparison.

PDF documents were the only document format with multiple logical pages on each technical page. Multiple logical pages were present in 13 documents, or 21% of the PDF document mass. Fully 38% of the multi-page documents consisted of two or six logical pages, while 23% consisted of four pages. The average number of logical pages per technical page was highest for documents with eight technical pages. Each of these technical pages consisted of four logical pages, making the document a 32-logical page document. The dataset did not include any documents with a multi-page facility for documents with 15 or more technical pages.

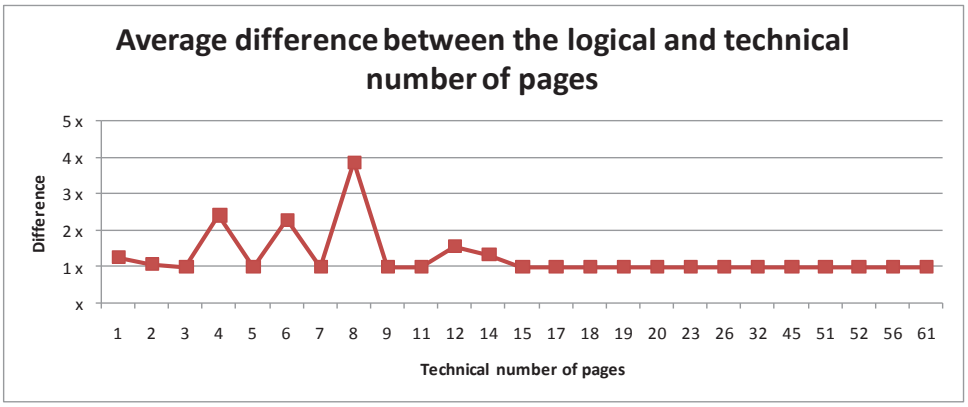


Figure 52: Difference between the logical and technical number of pages

The PDF documents did not include metadata on the number of logical pages or pages from the original converted document. Metadata harvesting is therefore unable to create metadata that indicate the logical number of pages. However, metadata extraction can

be performed: All the dataset's PDF documents with a multi-page facility proved to include a black frame around each slide, as shown in Figure 46 (p. 118). The number of these frames can be used to establish the number of logical pages in the document. Such a task can be undertaken either by analysing the document code directly or by using a content presentation application to recreate the visual appearance of the document before extraction efforts is undertaken.

Word documents

Technical Errors: Faulty "Pages" metadata

An initial analysis indicated that 45% of Word documents contained false "Pages" element entities. All documents with entities indicating more than one page were correct entities. However, documents with the entity "1" contained a false entity 82% (!) of the time. The degree of this error rate was far higher than expected.

Because of the unexpectedly high error rate, it was decided to reanalyse Word documents. A new selection of 100 random Word documents (DOC & DOT) was retrieved from the dataset. Ninety-seven had a "Pages" element with the entity "1" (one page). This element was correct for 31 of 97 documents. The entities indicated that no documents contained more than two pages. Forty documents contained more than two pages. This gives an overall error rate of 66%. The error rate for documents with more than one visual page reached 95.7% (!). However, all documents with an entity of more than one were in line with the visual observations. The entity "1" should hence be regarded as a default value for Word documents, which may or may not reflect the visual characteristics of the document. The number of pages presented through the normal user interface provided by MS Word applications is thus not useful in updating document metadata.

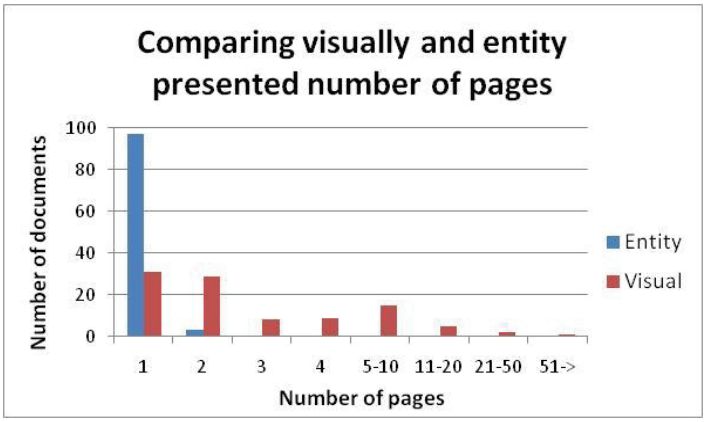


Figure 53: Comparing the "Pages" element with the visually correct number of pages

Extracting the visual number of pages

An analysis of Word 97-2003 (DOC) and Word 2007 (DOCX) document formats was undertaken in order to find alternative data sources for harvesting or extracting a correct "Pages" element. These document formats do not systematically include data about the visual characteristics of the documents. Manually created page breaks can be extracted, although other indications of where one page begins and ends are not stored as part of the document code. No documents in the dataset used this functionality.

The version of the MS Office Word 2007 application used in this research employs the visually presented number of pages as a metadata entity, and thus performs differently than the application versions that were used to create the documents in the dataset. It cannot be concluded that this is a new functionality since it has been observed previously, though not consistently. However, it does make it possible to experiment with documents as if this functionality was consistently present. For longer documents, the number of pages can be seen as a count in the lower left corner of the screen when a document is opened (tested on MS Word XP and MS Word Office 2007). If the document is not fully rendered when it is saved as a new document, the resulting metadata are wrong. For example, a document from the dataset consists of 87 pages and takes a few seconds to open on the computer used for this research, but when the same document was saved, , the number of "Pages" element entities totalled "12," "50" and "87," depending on when the document was stored. False entities can thus be generated if the document is not fully parsed. When importing or copying new content into a document, the documents are fully rendered before the saving process is executed. The MS Word applications base their efforts on characters and words. The "Characters" and "Words" elements are therefore constantly kept updated whenever the document is saved.

The extraction of correct "Pages" elements requires the use of a content presentation application in order to interpret the document code. It is essential that this application be able to interpret all content contained in the document code and that the document is fully rendered before any analysis of the document's visual appearance takes place. This procedure basically performs a virtual print-out of the documents using these characteristics as document metadata.

PowerPoint documents

All PowerPoint documents were found to have "Slides" elements with correct entities. However, it was noted that one of the documents also contained a "Pages" element. Further analysis of the final dataset revealed that the "Pages" element was present in 9.3%, or in 44 documents from the final dataset of PowerPoint documents. All these documents also contained a "Slide" element. The "Pages" element was not expected to be part of the metadata for PowerPoint documents as it is not a part of the document format's metadata schema. To determine what caused this to occur, all PowerPoint documents with a "Pages" element were analysed.

None of the "Slide" and "Pages" elements was identical. There was no obvious relation between the two elements: The "Slide" element varied from being 44 times higher than "Pages," to being 42 times lower than the "Pages" entity. All incidents did have one

thing in common: The “Slide” element was always the correct element. The “Pages” element did not contain the correct number of slides for any document. Further analysis revealed that all the documents involved were created using the application “PowerPoint 4.0” (anno 1994). All documents created by this application contained a “Pages” element with false entities. This research has concluded that this application generates false metadata “Pages” elements and entities. This shows that content creation software can generate metadata that violate the metadata schema of the document format and demonstrates the need to be familiar with the document format, its metadata schema and its practical usage before undertaking AMG efforts.

Summary

Chapter 4.3.1 has shown that *some* content creation software applications generate false entities, even including elements that are not present in the document format’s metadata schema. This has demonstrated the need for caution when using embedded metadata as a basis for document metadata descriptions.

All PDF documents contained a “Pages” element. The limited numbers of technical errors were all caused by documents with security restrictions, which denied access to the documents’ embedded metadata. PDF documents are frequently used to publish slideshows, in which multiple slides are commonly presented on each PDF page. This can cause a mismatch in the documents’ logical and technical number of pages. One-fifth of the PDF documents contained multiple logical pages. Multiple sub-documents can be included in each PDF document. The content of sub-documents was not presented through the master document’s metadata. The presence of sub-documents can cause logical and technical page errors. However, no such documents were found in our dataset.

All Word documents contained “Pages,” “Words” and “Characters” elements. The “Pages” and “Words” elements were presented in the MS Word application’s graphical user interface, although their entities were not necessarily equal to the entities that were used as embedded metadata. Technical “Pages” errors were found in two-thirds of the documents. Ninety-six percent of the Word documents with embedded metadata that indicated one document page contained false metadata. The number of document characters was inconsistently counted. Footers, footnotes, endnotes and headers were consistently not counted, resulting in too few records being recorded. However, too many characters were also counted. Similar observations were made regarding the number of words, which varied in an inconsistent manner.

PowerPoint documents should contain the “Slides” element instead of “Pages.” All “Slides” elements were visually correct. All PowerPoint documents created with the application “MS PowerPoint 4.0” contained both “Slides” and a false “Pages” element. This shows that common applications can generate metadata that violate the metadata schema of the document format, and demonstrates the need to be familiar with the document format, its metadata schema and its practical usage before undertaking AMG efforts. The “Words” element entities were on average lower than the visually present entities. These results were influenced by the content of imported content, such as illustrations, graphs and tables, which were not counted.

Word and PDF documents do not contain page break information (aside from manually created page breaks). These document types need to be fully parsed before the number of technical pages can be visually determined. The number of logical document pages can be determined by counting the number of logical characteristic page frames in the document. The number of words and characters can be determined by extracting this type of content, which is visible in the documents, and by counting the number of records.

This chapter has presented the value of combining use of the document code directly and use of content presentation applications to recreate visual appearance characteristics that are not explicitly stored as part of the document code.

4.3.2 The "Creator" element

The "Creator" element can provide important information about the origin of the document and can be regarded as providing quality information about the intellectual content of the document. This element should contain an entity with a single or multiple creator names, a group or organization name. A preferred person name consists of at least a given name and a surname. A person name can be formatted in a multitude of ways, e.g. by including abbreviations, middle names and the sequence of names as presented. Organization and group names can also be formatted in a multitude of ways. Due to different formatting of creator entities, Boolean comparisons are not sufficient to determine if candidate entities or other data sources are identical. Manual evaluation was therefore needed in this analysis to determine if entities were in fact the same creator(s).

The dataset

This analysis is based on PDF, Word (DOC & DOT) and PowerPoint (PPT & PPS) document formats. These document formats represented 91% of all the published stand-alone documents from the LMS used in this research. They support the inclusion of embedded metadata and formatted sub-sections of the document and can contain visible creator information. All the documents also have a full person name of the person who published the document to the LMS. From the final dataset, 100 PDF documents, 100 Word documents and 100 PowerPoint documents were selected at random for analysis.

Presence of visible creator information

Visual data to verify element content were present in only a limited way, which increased uncertainties and the ability to draw conclusions regarding the embedded metadata and the extracted metadata. Only 9% of Word documents contained such information, while 27% of PDF and 44% of the PowerPoint documents contained visible information. The scarcity of visible creator information has two consequences:

1. It can be impossible to evaluate the correctness of candidate entities based on AMG efforts.
2. AMG efforts based on visible characteristics need to be extremely careful not to generate entities for documents without visible creator information.

Due to these issues, any AMG efforts based on visual characteristics would result in entities of very low semantic quality. This research continued by analysing available data sources.

Harvesting creator metadata

Word and PowerPoint documents can contain "Author" and "Last author" elements. PDF documents can contain a general "Author" element and an Extensible Metadata Platform (XMP) section with "DC. Creator" and "XAP. Author" elements. Such elements have been harvested in related work [63]. An analysis of the dataset illustrated issues with entities from the XMP section:

- Additional characters were included to indicate the start and end of brackets: "\(" and "\)."
- Different characters were extracted: "" (blank) instead of "-" (line).
- The Norwegian character "ø" was replaced by "." (period).

All the "Creator" elements in the XMP section were present in the general element section as well. The general elements did not show these kinds of character errors. Hence, the general and XMP elements, which should have been synonymous with identical entities, do not have identical entities. These errors could not be traced back to a "faulty" application: The content creator software applications were commonly used with correct results. It is evident that there are issues regarding the content of the information placed in the XMP section of PDF documents. As a result, this research focused subsequent efforts on the general elements.

Table 17: Creator metadata from PDF, Word and PowerPoint documents

	PDF	Word	PowerPoint
Contain full author or organization name	11%	30%	38%
Visibly verifiable correct	4%	3%	6%
Not visibly verifiable correct	7%	27%	32%
Contain partial author or organization name	61%	36%	34%
Visibly verifiable correct	9%	2%	13%
Not visibly verifiable correct	52%	34%	21%
No results	20%	1%	7%
Verified false entities	8%	33%	18%

Author or organization names were contained in the metadata from 72% of the PDF documents contained, although only 11% of the metadata elements could be visibly verified to be correct. Eight percent of the documents contained false metadata, mainly as commercial content for online converting services. Sixty-six percent of the Word documents contained author or organization names in their metadata, but only 5% (!) of the metadata elements could be visibly verified as correct. A third of the entities could be verified as false with values such as "standard user" and "test." Seventy-two percent of the PowerPoint documents contained author or organization names in their metadata,

although only 19% of the metadata elements could be visibly verified as correct. Eighteen percent of the entities were verified as false.

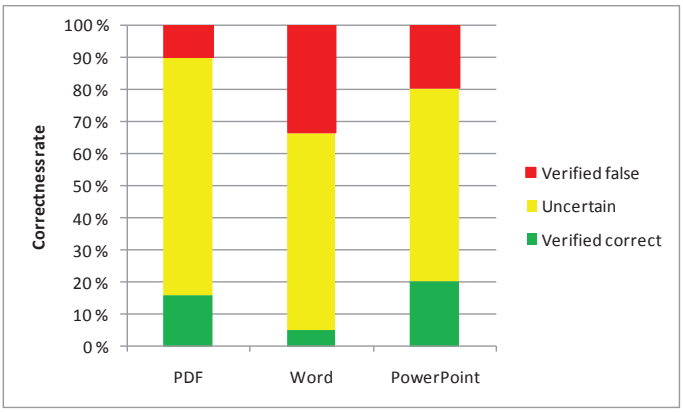


Figure 54: Verified correctness of embedded creator metadata

Figure 54 shows that 75% of PDF, 62% of Word and 60% of PowerPoint documents cannot be verified as either correct or false. Due to the lack of visible information against which to compare the embedded metadata, there is high uncertainty regarding the correctness of the harvested entities.

Table 18: Verifiable correct and false embedded creator elements

	PDF	Word	PowerPoint
Verified correct	16%	5%	20%
Uncertain	74%	62%	60%
Verified false	10%	33%	19%

Extraction using visual characteristics

Extraction based on visual characteristics has been performed by a number of researchers [49, 56, 85, 91, 122]. The current research has attempted using AMG based on visual characteristics in order to generate "Creator" element entities in a selected dataset. Table 19 shows that using the first line of text or the content with largest font does not generate "Creator" entities. These approaches are also more commonly used to generate "Title" elements. If the title can be correctly identified, then the likelihood of generating "Creator" metadata elements increases, although it is still low due to the limited number of documents with visible creator information. In all cases where visible creator information is not present, false entities are generated.

Table 19: Algorithms for generating "Creator" entities based on visual characteristics

	First line		Largest font		Located under the title	
	Correct	False	Correct	False	Correct	False
PDF	3%	97%	1%	99%	12%	88%
Word	1%	99%	0%	100%	4%	96%
PowerPoint	0%	100%	0%	100%	20%	80%

Extraction based on the document code

Word and PowerPoint documents can contain style tags that present the formatting used for specific sections of the document. No documents contained the "Author" or "Creator" style tags. Later versions of the Adobe PDF document format also support inclusion of style tags. This can allow retrieval of style formatted content from the original documents after conversion to PDF [109]. Six PDF documents contained format tags, though these referred to other content (descriptions of images).

Half of the PowerPoint documents contained "Sub-title" style tags. Sixty-eight percent of all visible creator information was found within this section. These sections were visually formatted in a variety of ways and contained a range of different data, such as subtitles, dates, course descriptions and creator information in a multitude of different orders. Creator information was included in 60% of the "Sub-title" sections. Eight percent of the "Sub-title" sections contained only creator information. The variety in regards to content types and visual formatting makes extraction efforts from this section reliant upon identification of user and organization names, among other text. This is a technology that has yet to be developed.

Table 20: Formatting information available from PDF, Word and PowerPoint documents

	Adobe PDF	MS Word	MS PowerPoint
Contain "Creator" or "Author" formatting	0%	0%	0%
Contained "Sub-title" formatting	0%	0%	50%
Section included creator info. only	0%	0%	8%
Section included creator info.	0%	0%	52%
Section did not include creator info	0%	0%	40%

Using the LMS publisher data as data source for the "Creator" element

An alternative to harvesting or extracting of creator metadata from stand-alone documents could be harvesting context publisher data from the LMS. Such an approach can generate valid entities for individual publishers. False entities would be generated for groups and organizations. Using an external data source for creator information has been performed by Greenberg [60] and Jerkins et al. [82].

Due to the limited number of publishers that are allowed access to the case LMS (only course lecturers), validation can be performed even though limited user information is available from the stand-alone documents. This research compared user profile names in the LMS against the embedded metadata. Positive results were obtained when entities that were related to the course authors were collected. For example, the harvested entity "Lars" would register as a positive match if the document was published by a "Lars" when no other "Lars'" could have made the publication. A match is considered positive if the publisher is included in the list of visible authors. This resulted in the correctness rates presented in Figure 55 and Table 21, which show a rate of 34% for PDF documents, 74% for Word and 55% for PowerPoint documents. This research also confirmed that the LMS publisher was not the document creator for 28% of the PDF, 7% of the Word and 35% of the PowerPoint documents.

Table 21: Verifiable publisher as document creator

	PDF	Word	PowerPoint
Verified correct	34%	74%	55%
Uncertain	38%	19%	10%
Verified false	28%	7%	35%

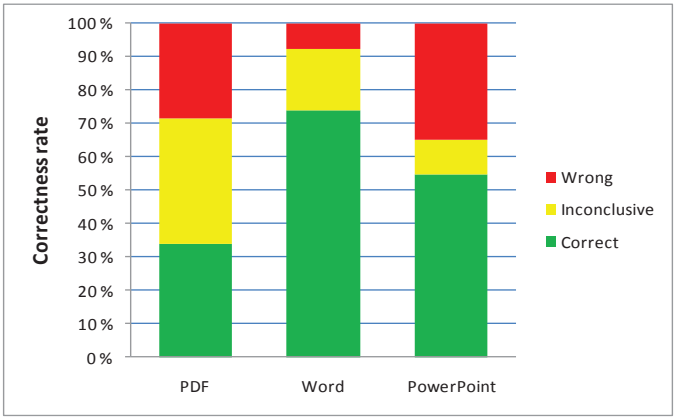


Figure 55: Verifiable publisher is document creator

This research also evaluated the correctness rate when a correct entity needed to contain the author names of *all* document creators, presented in Table 22. There are no differences regarding correctness for PDF and Word documents. This confirms that most Word documents are published by the document creator. Multi-creator Word documents are not commonly published. Rather, such documents are converted to PDF before being published. PowerPoint documents created by multiple persons are published in their original document format. Hence the correctness and false rates are

affected by the different requirements for verification of correct results; see Table 21 and Table 22.

Table 22: Stricter verification of publisher, including multiple authors

	PDF	Word	PowerPoint
Verified correct	34%	74%	47%
Uncertain	38%	19%	10%
Verified false	28%	7%	43%

Summary

This chapter presented the generation of "Creator" element entities. This analysis has demonstrated the challenge of not having a validated correct data source against which to compare the embedded and extractable data results. As a result, there are large uncertainties as to whether the generated entities are correct or false. This is due to:

- Content creation software that generates entities of low or very low semantic quality.
- Extraction based on visual characteristics which generates high quantities of false entities due to the use of data sources that do not contain the desired content.

This research has found that there is a potential for generating creator metadata based on creator style tags present in the document code. This approach would only generate entities when the desired content is present. However, due to the lack of practical use of document templates and use by document creators, this approach does not generate any entities for this dataset. The potential of this approach could therefore not be explored.

These harvesting and extraction efforts offer vastly contrasting results from the system controlled environment, where all documents were automatically given a valid creator element, as described in Chapter 4.1.2. Neither the consistency of information nor the correctness of the specific types of data available from stand-alone documents is comparable to the documents in the system controlled environment.

4.3.3 The "Title" element

This chapter analyses the embedded "Title" entities for common document formats and AMG approaches for generating such entities. Current research on AMG algorithms for generating the "Title" element is based on harvesting and extraction that rely on rules that use visual characteristics, as described in Chapter 2.3.5. This chapter presents a special focus on using the document code as the basis for extraction efforts. The PDF document code proved not to include content relevant for this analysis. Subsequent efforts were therefore focused on Word and PowerPoint documents. The documents were lossless converted to their respective Open XML document formats. Subchapters 1 and 2 in Chapter 4.3.3 present the baseline approaches and their results on this diverse dataset. Subchapters 3 to 9 in Chapter 4.3.3 present use of the document code as the basis for AMG efforts without and in combination with other AMG approaches.

The dataset and baseline experiments

The document code can contribute with data regarding non-visual content of each document. Such data are found in Word and PowerPoint documents, though they are discarded when original documents are converted to PDF. Only Word and PowerPoint documents were included in this analysis. The Word and PowerPoint document collection consisted of 974 documents, or close to 36% of the final dataset of stand-alone documents. From these documents, 100 Word documents and 100 PowerPoint documents were selected at random. Two corrupted PowerPoint documents were among the dataset. These were removed, leaving 98 PowerPoint documents in the dataset. The documents were converted to their respective Open XML document formats using the lossless converting functionality of the MS Office 2007 application suite. The converted documents were unzipped (extracted) and analysed as XML-based document code. The retrieved documents had a diverse visual appearance, ranging from being based on predefined official administrative templates created by university employees, to documents without any apparent structure created by students on private computers. Figure 57 and Figure 63 present two of the Word documents analysed. Figure 35, Figure 61, Figure 64 and Figure 67 present the PowerPoint documents that were analysed. This research conducted initial AMG efforts in generating baseline results based on the efforts of related work (see Chapter 2.3.5):

- **File name:** Obtained from the file system [14].
- **Embedded metadata:** Harvested from the document [57, 63, 82, 121, 123, 135].
- **First line:** Extracted from the first visible line of text [63].
- **Largest font:** Extracted the text section on the first page based on the largest font size [56, 57].

The approach of using the first line and largest font requires using content presentation applications for the recognition of visual characteristics. The first line approach uses these visual characteristics to gather the document's first visible line of text. The largest font approach requires using a set of weighted rules to evaluate the visual characteristics of the document. In related work, these rules were adapted to the specific dataset at hand, which was a dataset with documents sharing key visual characteristics aspects. With the diversity found in the visual characteristics of this dataset, such case-specific rules are not suitable. Instead, this dataset requires the use of rules based on the more general characteristics of a document title. The rules used for recognizing visual titles are presented in Table 23. Filtering of content has not been included in this effort due to the case-specific adaptations such an approach would require. Results from such efforts would therefore not be generally valid.

Table 23: Rule set for the largest font AMG baseline approach

- **Main rule:** Collect all content presented in the largest font
  - **Sub-rule 1:** Avoid the document header section.
  - **Sub-rule 2:** If all content has identical font, when the first line of text is used
  - **Sub-rule 3:** Prioritize collection of content with CAPITAL letters, then **bold**, underlined and lastly *italic* text.
- **Word document specific:**
  - Content must be placed on the top two-thirds of the first page
- **PowerPoint document specific:**
  - Content can be placed anywhere on the first slide
  - If no title were collectable from the first slide, then a title can be collected from the second slide

The results of the baseline efforts were categorized as correct, partly correct, no results and false results:

- **Correct:** The generated entity was identical or nearly identical to the visible title. Small variations, such as spaces that had been removed between words, were accepted.
- **Partly correct:** The generated entity was either partly correct or larger differences were present.
- **No results:** No content was generated by the algorithm. This can be the result of documents without embedded metadata or documents without text-based content.
- **False results:** The generated entity does not result in a representative "Title" element.

Baseline results

The results of the baseline experiments confirmed previous expectations:

- The file name tends to resemble the visible title, although the file name is frequently used to display additional types of data (such as dates and course code) in addition to a shortened title.
- The embedded metadata are strongly influenced by content automatically used as the title, as further explained in Chapter 4.3.3 subchapter 4.
- The first line approach frequently collects content from the document header, such as course codes, author names, dates and the number of pages. PowerPoint documents are affected by titles in large letters resulting in title information spread over multiple lines.
- Due to the similarities in visual presentation of a document title, the rule-based approach using visual characteristics performs much better than the other algorithms. However, there were a number of false results due to the collection of incorrect content, especially course information and person names.

Table 24: Results of baseline AMG Title algorithms: Word documents

	Correct	Partly collected	No result	False content
File name	40%	45%	0%	15%
Embedded metadata	27%	29%	8%	36%
First line	38%	15%	1%	46%
Largest font	69%	8%	1%	22%

Table 25: Results of baseline AMG Title algorithms: PowerPoint documents

	Correct	Partly collected	No result	False content
File name	21%	52%	0%	27%
Embedded metadata	28%	10%	0%	62%
First line	37%	34%	2%	28%
Largest font	76%	14%	2%	8%

The baseline results show that using the content with the largest font generated the most correct entities. The embedded metadata was strongly influenced by being automatically generated the first time the document was stored, and hence was not updated as the document evolved during the creation process. The first line algorithm frequently collected the document header section from page tops.

Content available from the document code

The original Word and PowerPoint documents can be lossless converted to Open XML document formats, which enables full access to all content of the document code as XML code. Open XML documents are zip archives containing standardized, structured content regardless of the document content. There are dedicated XML files for the header and footer sections. As a result, these sections can be avoided entirely. By analysing the content of the main document XML files for Word and PowerPoint documents, it is possible to analyse the main document content without the need for visual interpretations regarding font name and size, placements and section content. This chapter presents the files that are usable for AMG efforts in generating “Title” elements, and the types of data they contain.

Harvesting the embedded metadata “Title” element

The embedded “Title” element can be retrieved from the “Core.xml”-file located in the “docProps” folder of Word and PowerPoint Open XML documents.

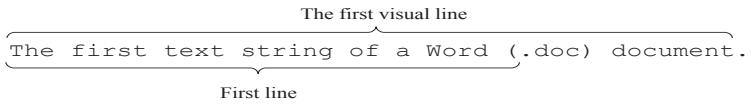


Figure 56: Distinction between the first visual line and the first line recorded

These elements are automatically populated with content generated by the content creation software. The MS Word applications used the first line of text in the document as the "Title" metadata element if no embedded entity was present. The MS definition of "the first line" is all characters until a line feed or period mark is encountered. This text line differs from the first visual line when a sentence covers multiple lines or if a period is present as part of the sentence, as is shown in Figure 56. MS PowerPoint applications automatically populate the "Title" element with content from the template section formatted as the "Title." This is only performed if the metadata "Title" element is without an embedded entity. Many PowerPoint templates have been observed to contain default "Title" entities, such as "No slide title" and "Slide 1." If the MS Word or PowerPoint applications find that there is an entity in the "Title" element, then the "Title" element remains unchanged until it is manually updated by the user. As documents are reused and re-titled, the metadata title element remains unchanged, and hence becomes false.

The MS Office 2007 applications do not automatically generate "Title" elements. Microsoft sees automatic generation of the "Title" element as a potential security issue because people are generally unaware of these automatically generated entities.

Gaining access to the main document content of Word documents

The procedures for gaining access to the main document content of Word and PowerPoint Open XML documents are not identical. This chapter presents techniques for accessing the principal document information for Word documents, while Chapter 4.3.3 subchapter 6 presents the same information for PowerPoint documents.

The document body of Word documents is accessible from the sequential "document.xml" file. The content listed at the beginning of the file is then presented at the beginning of the visual document. There are dedicated XML files for the header and footer sections. As a result, these sections can be avoided entirely. The "document.xml" file gives access to the document formatting, such as the user-specified Title and Heading sections. The section names from the document template that is used are visible in the document code.

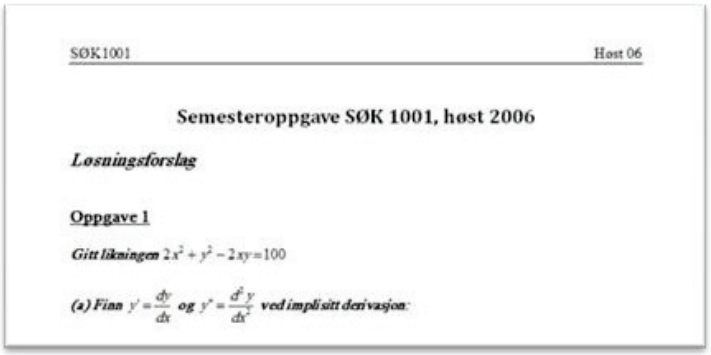


Figure 57: Example of a Word document with a visible title

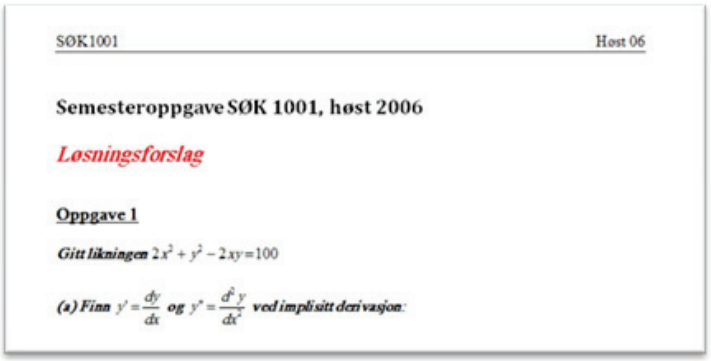


Figure 58: Example of a Word document with alternative visual presentation

Figure 57 shows an example of a document where there is a visible title. The title can be identified based on its placement in the upper part of the document, its bold, centred letters and large font. AMG rules based on visual characteristics can identify this title content, although even small formatting differences can confuse algorithms based on visual characteristics.

Figure 58 show the same document, but where the title is aligned to the left, and the sub-title has been increased in size and font colour. An analysis of the document code shows the actual formatting of the document. Figure 60 show the content of the main document's code, in which the style formatting tags of each content section are presented. The title can then be located by looking at the content of specific sections that are known to contain the desired content.

```
- <w:p w:rsidR="00DF5C95" w:rsidRDefault="00DF5C95" w:rsidP="004A6CFE">
  - <w:pPr>
      <w:pStyle w:val="Header" />
    - <w:pBdr>
        <w:bottom w:val="single" w:sz="4" w:space="1" w:color="auto" />
      </w:pBdr>
    </w:pPr>
  - <w:r>
      <w:t>SØK1001</w:t>
    </w:r>
  - <w:r>
      <w:tab />
    </w:r>
  - <w:r>
      <w:tab />
      <w:t>Høst 06</w:t>
    </w:r>
  </w:p>
</w:hdr>
```

Figure 59: Heading of the example document stored as a separate XML-file

```
- <w:body>
  - <w:p w:rsidR="009B6E37" w:rsidRPr="00CE45BC" w:rsidRDefault="009B6E37" w:rsidP="00D025E0">
    - <w:pPr>
        <w:pStyle w:val="Title" />
      </w:pPr>
    - <w:r w:rsidRPr="00CE45BC">
        <w:t>Semesteroppgave SØK 1001, høst 2006</w:t>
      </w:r>
    </w:p>
    <w:p w:rsidR="009B6E37" w:rsidRDefault="009B6E37" w:rsidP="009B6E37" />
  - <w:p w:rsidR="009B6E37" w:rsidRPr="004A6CFE" w:rsidRDefault="009B6E37" w:rsidP="00D025E0">
    + <w:pPr>
    - <w:r w:rsidRPr="004A6CFE">
      + <w:rPr>
        <w:t xml:space="preserve">Løsningsforslag</w:t>
      </w:r>
    </w:p>
    <w:p w:rsidR="009B6E37" w:rsidRDefault="009B6E37" w:rsidP="009B6E37" />
    <w:p w:rsidR="009B6E37" w:rsidRDefault="009B6E37" w:rsidP="009B6E37" />
  - <w:p w:rsidR="009B6E37" w:rsidRPr="00BC2515" w:rsidRDefault="009B6E37" w:rsidP="00D025E0">
    + <w:pPr>
    - <w:r w:rsidRPr="00BC2515">
      + <w:rPr>
        <w:t>Oppgave 1</w:t>
      </w:r>
    </w:p>
```

Figure 60: The Open XML document code of the example document.

In Figure 60 the visually present style tagged title was formatted as the “Title.” Identifying content in this way can avoid the need for other rules for locating the desired content. The document code only contains format content names of content formats used in the document. Figure 58 includes a header section. The content of this section is

placed in a separate file, called the "header1.xml" file, shown in Figure 59. This section can be avoided in the analysis if the main document XML file, the "document.xml" file, is used as data source. This file presents the main document content along with references to the formatting used. The actual formatting of each document section is mainly located in the "styles.xml" file, although this content can also be obtained directly from the main document. Using these data sources shows visual characteristics based on document facts, such as the font name, font size and colour, and whether the font is italic, bold, underlined, alignment, etc. This enables precise determination of the document content without the need for interpreting the visual content of the document.

Table 26: Formatting of the first three text sections of the Word document example

	Section 1	Section 2	Section 3
Content	Semesteroppgave SØK 1001, host 2006	Løsningsforslag	Oppgave 1
Line number	1	3	6
Style tagged name	Title		
Section format ID	00CE45BC	004A6CFE	00BC2515
Font name	Cambria	Times New Roman	Times New Roman
Font size	16	17	12
Bold	Yes	Yes	Yes
Italic		Yes	
Underline			Yes
Colour	Automatic (black)	Red	Automatic (black)
Alignment	Left	Left	Left

Gaining access to the main document content of PowerPoint documents

The structure of PowerPoint Open XML documents differs from Word documents, although the principles are similar. These documents consist of a compressed archive with dedicated XML files specifying specific content in the document. It is therefore possible to generate tables for PowerPoint document content formatting as shown in Table 26.

Each PowerPoint slide corresponds to a dedicated "slide.xml" file. This makes it possible to work on a specific slide. The common slide template content is stored in a separate file, similar to Word document headers. PowerPoint documents are not sequential. Instead, each object on the slide (e.g. text, multimedia content) is given X and Y coordinates for horizontal and vertical placement. Locating the text box that is visually on the top of the page requires a comparison of all the text box coordinates on the specific slide.

Text can be formatted as a specific format style in Word documents. In contrast, PowerPoint content is given the same format for an entire text box. Due to this "boxing" of content, all content of a given section is placed in the same section of the document code. This enables more efficient collection of complete text sections, even if multiple text sections are located on the same page or if text crosses multiple visible lines. Content within each text box can be formatted individually. This enables rules that are based on visual characteristics to make distinctions between different document contents.

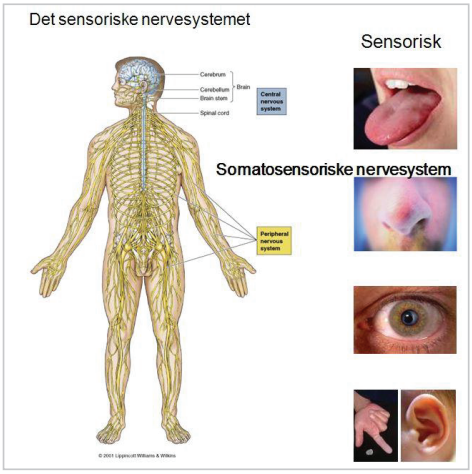


Figure 61: PowerPoint slide with text boxes, images and groups of content

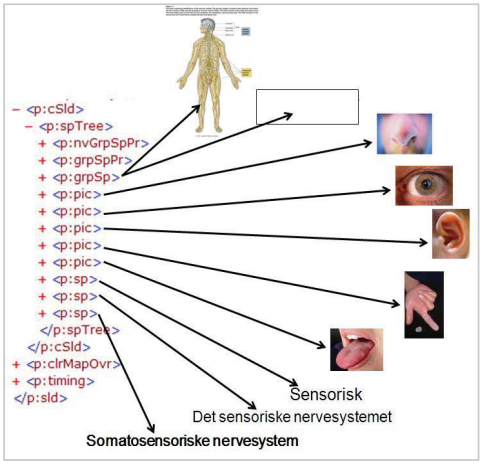
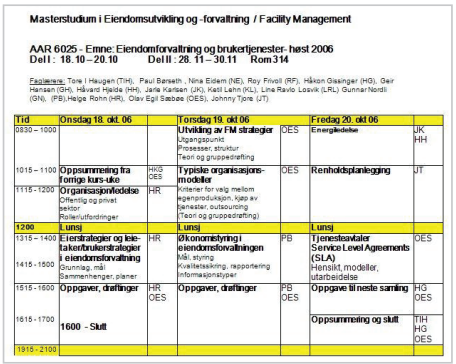


Figure 62: Placement of the slide content in the "slide.xml" file.



Masterstudium i Eiendomsutvikling og -forvaltning / Facility Management

AAR 6025 - Emne: Eiendomforvaltning og brukertjenester- høst 2006
Del I : 18.10 – 20.10 Del II : 28. 11 – 30.11 Rom 314

Fagteam: Tore I Haugen (TIH), Paul Barseth , Nina Eidem (NE), Roy Frivoll (RF), Håkon Glesinger (HG), Geir Hansen (GH), Håvard Hjelde (HH), Jarle Karlsen (JK), Kjell Lehn (KL), Line Ravlo Losvik (LRL) Gunnar Nordli (GN), (PB),Helge Rohn (HR), Olav Egil Sæbøe (OES), Johnny Tjore (JT)

Tid	Onsdag 18. okt. 06		Torsdag 19. okt 06		Fredag 20. okt 06	
0830 – 1000			Utvikling av FM strategier Utgangspunkt Prosesser, struktur Teori og gruppedrefting	OES	Energiledelse	JK HH
1015 – 1100	Oppsummering fra forrige kurs-uke	HKG OES	Typiske organisasjons-modeller	OES	Renholdsplanlegging	JT
1115 - 1200	Organisasjonsledelse Offentlig og privat sektor Rollerutfordringer	HR	Kriterier for valg mellom egenproduksjon, kjøp av tjenester, outsourcing (Teori og gruppedrefting)			
1200	Lunsj		Lunsj		Lunsj	
1315 – 1400 1415 - 1500	Eiendstrategier og leie-taker/brukerstrategier i eiendomsforvaltning Grunnlag, mål Sammenhenger, planer	HR	Økonomistyring i eiendomsforvaltningen Mål, styring Kvalitetssikring, rapportering Informasjonstyper	PB	Tjenesteavtaler Service Level Agreements (SLA) Hensikt, modeller, utarbeidelse	OES
1515 - 1600 1615 - 1700	Oppgaver, drøftinger 1600 - Slutt	HR OES	Oppgaver, drøftinger	PB OES	Oppgave til neste samling Oppsummering og slutt	HG OES TIH HG OES
1915 - 2100						

Figure 63: Word document with a spreadsheet and visual "Creator" element as course name

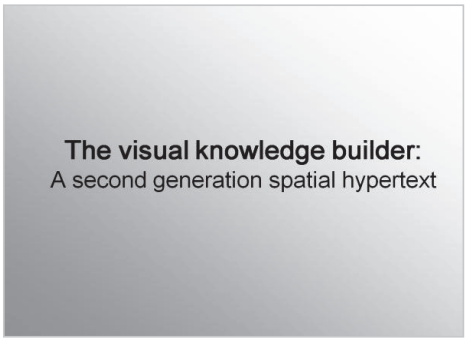


Figure 64: PowerPoint document with multiple types of content in a single text section

Figure 64 shows a document where all visible content is placed in a single text box formatted as the "Title." By using rules similar to those developed for rules based on visual characteristics, this text box can be classified into two content types:

- "Title" = "The visual knowledge builder:"
- "Sub-title" = "A second generation spatial hypertext"

Making such distinctions without having data that states that these elements are related can easily result in false results.

Sub-titles can also be generated by retrieving content that has been formatted in the "Sub-title" style. Figure 67 shows an example of a document template where the user has specified a title and a sub-title in different text sections. Figure 65 presents the template that was used. The identification of these text sections can be located in the XML file of the specified slide, presented in Figure 66.

Figure 65: Template with title and sub-title sections

Figure 66: XML document code for Figure 67

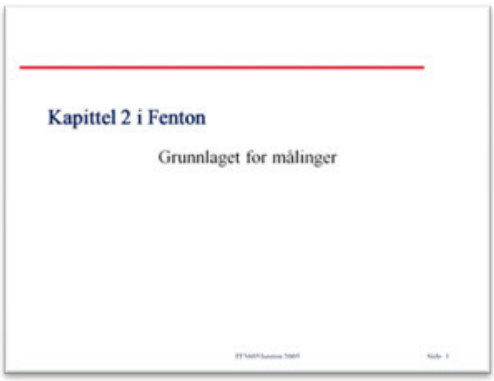


Figure 67: The template in Figure 65 in use

Results of using style tag formatting

The largest font approach could have achieved better results with the use of a custom case LMS content filter, although this would result in a local, case-specific algorithm solution. This research continues by presenting how the document code can be used to generate elements regardless of visual characteristics and in combination with other AMG approaches.

Table 27: Results of using style tag formatting

	Word	PowerPoint
Contains a "Title" section	3%	82%
Contains a "Title" section with a formatted "Sub-title"	1%	7%
Contains a dedicated "Sub-title" section	0%	38%

Only three Word documents contained "Title" style tags. Two of these documents used the style to format data other than the visible title.

In datasets where templates are more actively used, this approach has a great deal of potential. This can be seen in the results from the PowerPoint documents, where 82% contained "Title" style tags. The style formatted content contained representative titles in all cases. The sub-titles found in the "Title" style sections were valid sub-titles in the form of a continued title presentation. The content collected for "Sub-title" consisted of a continued title element, author name, date, and course and institute information. Two-thirds of the documents with a dedicated "Sub-title" used this section to present author information.

Combining AMG methods

The key property that allows the document code approach to be combined with other AMG methods is that it does not deliver a result when the desired content is not located. This enables it to be combined with other AMG methods. Our research demonstrated this by testing three different document code-based algorithms:

A. **Document code exclusively:** Generates "Titles" elements based exclusively on the document code. No other data sources or algorithms are used.
B. **Document code and largest font:** Extends algorithm A by evaluating if algorithm A provides an entity. If not, then the content with the largest font section is collected. These rules are based on rules that have been previously presented in Table 23, although adapted to this new dataset as presented in Table 28.
C. **Document code, largest font, context filter and alternative data sources:** Extends algorithm B by evaluating if algorithm B provides an entity after performing context data filtering (e.g. course codes and course descriptions). The largest font sub-algorithm can be executed twice if the first attempt results in a blank entity after filters have been applied. If no entity is generated, then the embedded metadata entity is harvested. If this entity is empty then the file name is used as the entity.

Table 28: Comparing rules based on visual characteristics and the document code

	Visual characteristics	Document code
Main rule	Collect all content presented in the largest font	Collect all content presented in the largest font. Font size collectable from the "styles.xml" file and the main document
Sub-rule 1	Avoid the document header section	Do not use the "header.xml" file
Sub-rule 2	If all content has identical font when the first line of text is used	If all content has identical font when the first line of text is used
Sub-rule 3	Prioritize collection of content with CAPITAL letters, then **bold**, underlined and lastly *italic* text	Prioritize collection of content with CAPITAL letters, then content formatted with **bold**, underlined and lastly *italic* text. These characteristics can be retrieved from the "styles.xml" file or from the main document: • Bold: <w:b /> • Underlined: <w:u /> • Italic: <w:I />
Word document specific	Content must be placed on the top two-thirds of the first page	-
PowerPoint document specific	Content can be placed anywhere on the first slide. If no title was collectable from the first slide, then a title can be collected from the second slide.	Content can be placed anywhere on "slide1.xml." If no title was collectable from the first slide, then a title can be collected from "slide2.xml."

Due to the lack of page information in the document code, the Word document-specific rule cannot be directly transferred to the document code approach. There are two approaches to solve this: A content presentation application could be used to interpret the document content, or a word counter could be implemented as an alternative sub-rule. An analysis of the dataset revealed that neither effort would have affected this dataset and the results. This effort has therefore been left out of subsequent models.

The algorithms start by converting binary Word and PowerPoint documents into Open XML. The Open XML document is extracted (unzipped) before the individual algorithms perform their tasks. Figure 68, Figure 69 and Figure 70 show the logical structure of algorithms A, B and C.

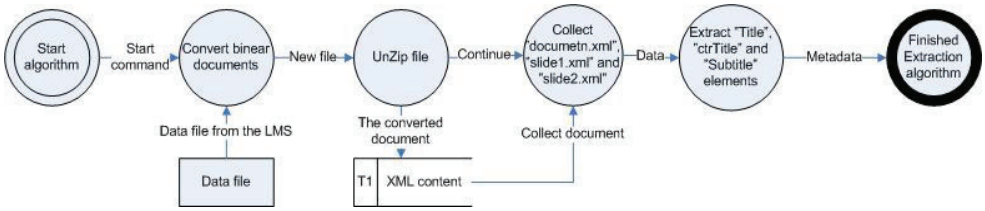


Figure 68: Logical structure of algorithm A

Algorithm A extracts the "Title" style tagged content, uses it directly as metadata and finishes execution.

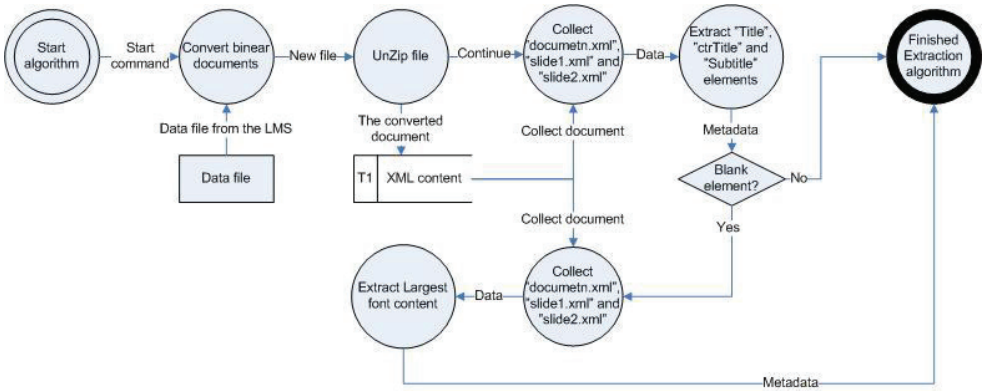


Figure 69: Logical structure of algorithm B

Algorithm B extracts the "Title" style tagged content. If the executed algorithm results in an entity (not blank), then these data are used as metadata and the algorithm finishes execution. If no result is generated, then the algorithm retrieves a dataset to which the rules based on visual characteristics are applied.

Algorithm C extends the previous algorithm by including filtering of unwanted content and the use of other alternative data sources. A filter (Filter nr 1) is added to exclude course data placed in the beginning of style tagged content. A loop has been included for the largest font sub-algorithm, allowing the algorithm to execute multiple times if filter nr 2 removes all content generated by the largest font sub-algorithm. The primary focus of the filter processes nr 1 and 2 is to exclude context information. These data can be collected from the individual course section of the LMS where the document was published. The filters were adapted to exclude the course code, the official course name (in Norwegian and English) and the institution name ("NTNU"), either abbreviated or not. If the filtered data is not blank, then this content is used as metadata. If the largest font sub-algorithm does not result in an entity (after filtering), then the embedded

metadata element is collected. The content of this sub-algorithm is filtered by filter nr 3 in order to exclude default entities, such as “Document1” and “No slide title.” If no content is generated after filtering, then the document’s file name is used as metadata.

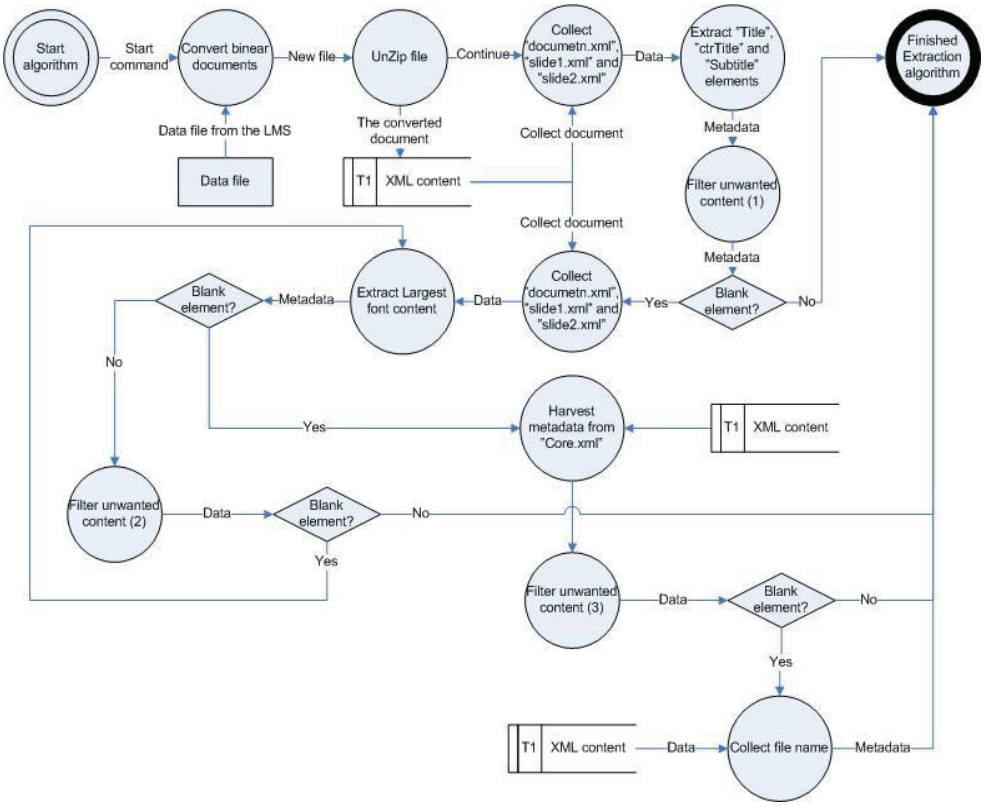


Figure 70: Logical structure of algorithm C

Results of the document code based efforts

The falsely labelled Word document appeared in the algorithm results, see Table 29. As these AMG efforts were constructed to demonstrate the possibilities of using the document code, these results have been accepted. In a real-world scenario, use of the “Title” style tags for Word documents would not be recommended for use if these tags are used in a way that is similar to what was observed in this dataset.

Algorithm B is not able to take advantage of the formatted titles in the Word documents. Instead, the false results of algorithm A are transferred to algorithm B,

reducing its correctness rate and increasing the percentage of false content records. The inclusion of context data filters in algorithm C resulted in the number of false records being reduced from 22 to 3. This resulted from the course and institution data being removed from the "Title" element. One document was given a title based on the file name, since neither the document body nor the embedded metadata contained text-based content. No filtering of default values was undertaken, hence filter nr 3 was not used.

Table 29: Results of advanced AMG Title algorithms: Word documents

	Correct	Partly collected	No result	False content
Algorithm A: Document code exclusively	0%	0%	98%	2%
Algorithm B: Document code and Largest font	71%	6%	1%	22%
Algorithm C: Document code, Largest font and filters	91%	6%	0%	3%

In this dataset the three documents that received a false entity based on the style tags would have received a correct entity based on their visual characteristics. Excluding use of the style tags for title generation would increase the correctness rates of algorithm B and C by two percentage points, while reducing the percentage of false content by two percentage points.

Table 30: Results of advanced AMG Title algorithms: PowerPoint documents

	Correct	Partly collected	No result	False content
Algorithm A: Document code exclusively	85%	0%	15%	0%
Algorithm B: Document code and largest font	94%	0%	3%	3%
Algorithm C: Document code, largest font and filters	97%	0%	0%	3%

Algorithm A takes advantage of the PowerPoint documents' style formatted "Title" content. All these formatted sections contain valid titles, either as the title, sub-title or a combination of both. The remaining AMG efforts from algorithm B were concentrated on the documents without a style formatted title. This resulted in one document being assigned a false entity while three documents were assigned a correct title.

The results from algorithm B were further improved by algorithm C. Desired content were not incorrectly filtered. One document was given a title based on the file name. There was no filtering of default values, hence filter nr 3 was not used.

Summary

In this dataset, the Meta tags for the style type title were seldom used in Word documents. In two of three cases these tags were used to format content other than the title. However, in regards to the PowerPoint documents, these style tags were extensively used and used correctly. This shows that the user's habits and the templates he or she uses strongly influence the document code and hence also the potential for automatic generation of metadata based on the document code.

The AMG efforts associated with algorithm B focused on documents for which there were no results from algorithm A. This resulted in a large portion of correct records, although with some errors. The inclusion of context data filters in algorithm C greatly reduced the number of false records. One document was given a title based on the file name, since neither the document body nor the embedded metadata contained text-based content. By excluding use of algorithm A for Word documents, the correctness rate would increase by two percentage points, reducing the number of false records by a similar amount.

Algorithm A employed the "Title" style tags that are frequently included in PowerPoint documents. All these sections contained valid titles. The remaining AMG efforts of algorithm B then concentrated on documents that did not have a style formatted title. This resulted in one document being given a false label while three documents received a correct title. Algorithm C gave titles based on the file name to documents without text-based content. No filtering of content was performed.

This research shows how the document code can be used as an informative data source to determine a document's visual title. Use of the document code allows implementation of precise rules based on visual characteristics, resulting in the collection of specially formatted text and complete text sections and lines. This has reduced the number of partly collected titles from 8% to 6% for Word documents and from 14% to 0% for PowerPoint documents. Using the document code has enabled the collection of preformatted sub-titles by retrieving content style tags that were formatted "Sub-title" or by combining the "Title" style tag formatted content with their visual characteristics.

The analysis showed that local filters need to be included to increase the correctness rate of the AMG algorithms. This research has shown that such filters do not need much local adaptation if a dedicated data source is used as the basis for excluding content. In this case, the correctness rate was increased by filtering official course information and institution information.

4.3.4 The "General. Language" element

This chapter presents the use of existing, automatically generated language tags for common document formats for AMG purposes. The document code can contain tags reflecting the language of the document's intellectual content. This can be used for the IEEE LOM's "General.Language" element [74] and the execution of AMG algorithms based on natural language in multi-lingual environments.

Purpose of efforts and the dataset

The IEEE LOM's "General. Language" element [74] relates to the language used for a document's intellectual content. It is not to be confused with the element "Educational. Language" element, which reflects the primary spoken or written language used by the intended user group. AMG algorithms based on natural language can be used to generate keywords and descriptions, and perform classification, as just a few examples. These efforts require knowledge of the language of the document's intellectual content. This knowledge is manually built into the efforts of related work by applying a default language. In a multi-lingual user environment, as is found in the case LMS, such an assumption would not hold as documents are published in a number of different natural languages. Current AMG algorithms based on natural language can hence not be executed win this user environment since the natural language of the intellectual content is undetermined. The qualitative analysis of Chapter 4.2.1 showed that none of the document formats that were published in the case LMS contained embedded language metadata.

The document code can contain language tags that indicate the language of the document's intellectual content. This allows for populating the "General. Language" element and the execution of AMG algorithms based on natural language. Language recognition is automatically performed by applications such as MS Word and MS PowerPoint on document text sections to enable spelling and grammar checks. These section-wise language descriptions are stored as language tags in their created Word and PowerPoint documents. Our research documented that language tags are discarded if the document is converted to a PDF. This research is hence focused on Word and PowerPoint documents. The language tags can be presented when Word and PowerPoint documents are lossless converted to their native Open XML document format. This chapter presents how these data sources can be used in order to generate metadata.

One hundred documents were selected at random, resulting in 60 Word and 40 PowerPoint documents. These documents were lossless converted to their native Open XML document format. The analysis was performed on the main document content of Word documents and on the first slide of PowerPoint documents.

Locating the language tags

An introduction to the content of Open XML documents from Word and PowerPoint documents is presented in Chapter 4.3.3 on page 138. Content creator software includes language tags in the main document ("document.xml") and in the document description file (a DTD-file), called "style.xml" in Word Open XML documents. PowerPoint Open XML documents do not contain this type of a document description file. Instead, a whole range of files can contain language tags, such as files for the template master, the header, footer and each individual slide. These language tags look like the examples below:

Example 1: <w:lang w:val="en-US">

Example 2: <a:rPr lang="nb-NO">

Language tags can be located in multiple places in a single document, allowing sections to be tagged with different language formats. Default language tags are assigned based on the language of the user interface in the content creator software used in creating the document. This research analysed the "document.xml" and "styles.xml" files from Word Open XML documents, and the first slide of PowerPoint Open XML documents retrieved from the "slides1.xml" file.

Case results: Word documents

The analysis revealed that the language tags located in the document body and the description file gave misleading impressions of the language of the document's intellectual content. This is because there are language tags in the document that are not in practical use. *All* the analysed documents contained "en-US" (US English) language tags, even though only 7.5% of these *used* these language tags. Extraction of all available language tags and using them to specify the language of the document's intellectual content will thus result in a low correctness rate. Extraction efforts need to be focused on tags that are in practical use. The extraction effort showed that all text sections were formatted with a single language tag. This allows the use of language-specific natural language AMG algorithms on individual sections formatted with a specific language tag. Both single and multi-lingual documents were found. As far as could be determined by this research, the language tags were placed correctly in accordance with the language of the document content[16]. These documents contained language tags indicating that their intellectual content was in:

- Norwegian ("nb-NO"): 42 documents
- British English ("en-GB"): 8 documents
- US English ("en-US"): 3 documents
- Danish ("da-DK"): 1 document

There was a clear majority of documents in Norwegian, followed by British English, US English and Danish.

Related research on AMG efforts that are based on natural language operates on the assumption that the document's intellectual content is in a single natural language. In a multi-lingual publishing environment documents with content in multiple natural languages can be present. The analysis revealed that six documents contained multi-lingual text sections. Distinguishing between content sections containing different natural languages enables extraction efforts based on the correct section language. Hence, each language section can be analysed separately. This approach avoids contamination of the dataset caused by content of other language(s) than the section language. The multi-lingual documents that were found were in:

[16] The data results (from both the Word and PowerPoint analysis) included references to "US English," "British English," "Australian English," "Greek" and "Brazilian Portuguese." This thesis has not conducted an analysis to determine if the dialects of the main languages of English, Greek and Portuguese match the language tags, as long as the language tags were representative of the main language.

- New Norwegian and English (“nn-NO” and “en-US”): 3 documents
- Norwegian and English (“nb-NO” and “en-US”): 2 documents
- English and German (“en-US” and “de-DE”): 1 document

Case results: PowerPoint documents

PowerPoint documents typically contain a limited number of complete sentences for which language recognition can be performed. Hence less data is commonly available with which to determine the language used in the document. This can result in less accurate language tags than for Word documents. All but one document contained language style tags that were in use. The exception only contained photographs and no text. In this case, the fact that no language was specified is regarded as the most correct.

Single language PowerPoint documents were found in Norwegian (17 documents), US English (5 documents) and British English (3 documents). One document used false language tags, when a few Norwegian keywords were included on the first slide of an US English slide show. This illustrates the difficulties posed by recognizing short language sections.

Thirty percent of the PowerPoint documents were correctly labelled as containing multi-lingual intellectual content. All these documents were formatted correctly. All these documents contained extensive text sections in the primary and secondary intellectual language. This has enabled the content creation software to correctly identify the intellectual language of each document section. These documents were in:

- Norwegian and US English (“nb-NO” and “en-US”): 9 documents
- Norwegian and Portuguese (“nb-NO” and “pt-BR”): 1 document
- Australian English and US English (“en-AU” and “en-US”): 1 document
- Greek and US English (“el-GR” and “en-US”): 1 document
- British English and Norwegian (“en-BR” and “nb-NO”): 1 document
- US English and Norwegian (“en-US” and “nb-NO”): 1 document
- US English and Swedish (“en-US” and “sv-SE”): 1 document

The primary intellectual language of these documents was correctly formatted with style tags. The language tags of the secondary intellectual language content were used to format number characters without any text. These language tags are not false, though this research regards these secondary language style tags of being misleading. Language style tags from document sections containing only numbers and symbols (not text) could be considered to not have been extracted.

Summary

This analysis has confirmed that the case LMS is a multi-lingual publishing environment. Documents were observed with intellectual content in Norwegian (“Bokmål”), US English, British English, Australian English, New Norwegian, German, Greek and Danish. Other analyses undertaken as a part of this research found documents in Canadian English, Swedish, French and Spanish.

All the Word and PowerPoint documents with intellectual content used language style tags. This research has shown that the language of a document's intellectual content can be determined correctly for nearly all Word and PowerPoint documents by extracting the language style tags that are in use. Using the document code and its language style tags enables identification and segmentation of documents into sections based on a specific natural language. This allows extraction efforts based on natural language to be executed in a multi-lingual user environment and for documents containing intellectual content in more than one natural language.

4.3.5 Qualitative Summary

The qualitative analysis has to an extent confirmed observations from the qualitative analysis: Embedded metadata from documents contain a high latent possibility for false entities. Without any central control over a document, the document's content is highly influenced by the local applications used to create it. These analyses have confirmed that automatically generated entities generated by document applications frequently contain false entities.

This research has demonstrated the effects that existing harvesting and extraction algorithms had on this dataset, which has shown that existing technologies are not optimal for gathering entities that reflect the documents. More accurate entities can be obtained by extracting specific document content and using this information in specific ways. These results also show how existing AMG technologies can be used with increased accuracy and reliability. This results in automatically generated metadata of higher semantic quality. The document code uses as presented in this chapter suggests the possibility of using AMG in environments where current AMG efforts would not perform due to the diversity in visible characteristics and natural language of the documents' intellectual content.

The document code in Word and PowerPoint documents show all content in a document. This data source can hence be used for harvesting and extraction efforts other than presented by this research. For example, tables can be extracted with all content in correct columns and cells even without visible borders, because all tables are assigned table formatting that can subsequently be retrieved. These data would be of value for research efforts like [97]. Tables of contents, which are automatically generated, are also recognizable based on their document code formatting. This data source could populate qualified Dublin Core "Description" elements. Recognition of tables and tables of contents are just two examples of additional opportunities that are possible based on an analysis of the document code.

5. Conclusion, contributions, objectives and future work

This section brings together the threads created in the previous sections in order to create the larger picture.

5.1 Conclusion

AMG algorithms base their efforts on systematic and consistent properties of the documents at hand in order to generate quality metadata in accordance with pre-defined metadata schema(s). AMG algorithms need to find common structures in which to base their efforts, even if the dataset is not homogenous. Recognition of the most correct and most desirable document properties is the basis for automatic generation of high quality metadata.

This research has documented that the document code can be used to automatically generate metadata of high quality even though the data source is not homogenous. Common, non-visual document formatting that can be obtained through the document code enables the generation of high quality metadata. This code is unique for each document format, although it is shared by all documents of the same document format version. The document code allows for the unique identification of all sub-sections of the documents and enables extraction from each formatted section individually, which in turn allows for the generation of a multitude of different metadata elements. AMG efforts based directly on the document code only generate results when the desired content is present, avoids interpretation of the document content and can provide other AMG algorithms document descriptions based on facts. These properties enable efficient combinations of AMG algorithms, allowing different harvesting and extraction algorithms to work together in order to generate the most desired results.

Extraction efforts based on the document code are vulnerable to having false content included by content creation software, by the template used and by the user. Such false content includes: (1) False content generated by the content creation software; (2) Extraction of falsely formatted sections (e.g. the ‘Title” formatted document section of Word documents consisting of author information); (3) Document descriptions that describe the template rather than the finished document; (4) Content included by the content creation software without the content being in use (e.g. language tags in Word and PowerPoint documents).

The ability to only generate entities when the desired content is present enables document code algorithms to be combined with other AMG efforts when these can generate more desired results. These entities are elements that cannot be obtained through the direct analysis of the document code (e.g. based on visual characteristics of Word documents) or when the document code systematically contains data that are not the desired results.

Using the document code as the basis for other extraction efforts enables direct access to the document content without contamination caused by content presentation applications. This ensures access to the most detailed and accurate document content

descriptions plus navigation possibilities within the document to formatted sections of interest for these algorithms. This in turn enables extraction efforts to be based on visual characteristics and the natural language approach, which provides a more desirable data source and hence better results than what has been previously possible.

The user environment in which a document is created has a strong influence on the resulting document. This is clearly apparent in the systematic and consistent properties of the documents:

- Documents from the system controlled environment have a pre-specified structure built on pre-defined templates. This ensures that each document section is pre-determined. The system controlled environment can enforce control of each section's value space, ensuring that only valid content defined by the document and metadata schemas are included. The system controlled environment enables AMG efforts to be based on known, systematic document characteristics.
- The uncertainties are extensively greater regarding the content of stand-alone documents. Different applications, application versions and templates contribute to the documents' content in a number of different ways. The user is given the intellectual freedom to decide how to use the applications, templates and user defined efforts. Finding the systematic and consistent properties of stand-alone documents can therefore require considerably more or diverse efforts.

AMG efforts based on the system controlled environment are based mainly on interpretation of the LMS document types. AMG efforts based on stand-alone documents require an understanding of how the documents are used by the document creators (users), what the user specifies and what is automatically generated based on templates and application specific AMG algorithms. This research has documented that such efforts can generate high quality metadata from stand-alone documents from a non-homogeneous dataset.

These efforts are based firstly on the recognition of the different properties of each stand-alone document format. Secondly, the document creator application version can give extensive consistency information regarding the elements that are being used and how these are used. Thirdly, template information can give further information that enables identification of the template used and the adaptation of the AMG algorithms to the specific template to maximize their ability to retrieve entities and maximize the quality of these entities. In addition, the document context can be used to generate extensive context-based metadata descriptions and increase the accuracy of document content based AMG algorithms. This research has presented how AMG efforts can be combined in order to generate high quality metadata from both a controlled and a user controlled document creation environment.

5.2 Major Contributions

This thesis has documented that significant amounts of metadata can be automatically generated from commonly used document formats based on stand-alone documents, or documents from a system controlled environment. However, the generation of *high quality* metadata requires the selective use of available data sources and specific

algorithms to maximize the potential of each data source. The major scientific contributions from this research have substantiated to the following conclusions:

- Efficient AMG efforts can be conducted on non-homogeneous documents and hence gaining usability of a single AMG algorithm regardless of subjects, language of intellectual content and the documents' visual characteristics.
- A vast majority of text documents are created by using the MS Office application suites to create MS Office document files. Efficient AMG efforts can be conducted directly on MS Office documents, and hence gain usability outside of the educational community, such as for home or business (intranet) usage.
- Using the document code as the basis for AMG efforts enables AMG algorithms to access, navigate and retrieve all of the content in a document, and to perform actions based on the facts presented in the document rather than the impressions created as a result of content presentation applications. This ensures that the user-specified, intellectual content, template information and content creator software, including formatting information, is available and undistorted regardless of the visual characteristics of the documents or the language of the metadata and intellectual content.
- Using the document code as the basis for AMG efforts enables the harvesting and extraction algorithms to be efficiently combined in order to maximize the quality and quantity potential of available algorithms.
- The user environment and the actions performed by the user, the content creation software and the template content all have a significant impact on the quality of the data sources that are available through the document code. Making the user aware of document properties and promoting the intended use of these can significantly increase the completeness and quality of data sources available for AMG algorithms.
- AMG is usable for generating extensive metadata descriptions based on document specific data and context data. This enable automatic generation of SCORM Learning Objects based on existing, published documents.

5.3 Reaching the Objectives

This research used a set of objectives as framework for the executed efforts introduced in Chapter 1.5. This chapter summarizes if the objectives of this thesis has been met or not.

RO1: Examine how commonly used content creation software (applications) use document code to store metadata, formatting data and intellectual content.

Results: *This thesis has examined documents regardless of their creator software. By using datasets retrieved from educational and business environments without filtering of "undesired documents", large amounts of documents have been analysed created using common software (applications). The research results are hence of higher value for "everyday use" than e.g. basing the research on a specific document collection or a specific document type.*

RO2: Document the kinds of metadata, formatting data and intellectual content that are contained in the document code of commonly used document formats.

Results: *This thesis has examined the document code of common file formats. By doing so, this thesis has documented a variety in quality in terms of accuracy and presence of metadata, formatting data and intellectual content which could be used for metadata Harvesting and/or Extraction efforts.*

RO3: Substantiate how document conversion between incompatible document formats influences the metadata, formatting data and intellectual content of the resulting document code.

Results: *This thesis has documented that a conversion process commonly affect visible and non-visible document content. This thesis has presented how such conversion processes results in initially harvestable metadata being excluded or replaced, that meta-tagged content sections lose their meta-tags, and visual appearance change. From an AMG perspective, converting documents include a significant danger of corrupting the data source. It is hence of benefit for AMG-algorithms to work with original data sources and hence necessary to understand the documents' original file format in order to make conversion between file formats unnecessary.*

RO4: Explore the possibilities for metadata extraction based on the document code and the consequences these efforts have on the quality of the generated metadata.

Results: *This thesis has performed extensive metadata extraction efforts based on the document code. By doing this, this thesis has gained direct access to more document content which the authors themselves have specified, giving the AMG algorithm extensive advantages in order to create high quality metadata. However, not all of the extractable content was applicable for creating high quality metadata due to template content and document content that is in conflict with the document template. AMG algorithms based on the document code could hence benefit from filtering and other types of generating metadata.*

RO5: Explore the possibility of using the document code in combination with or directly as the data source for other extraction efforts based on visual characteristics and natural language AMG technologies, without the need for content presentation applications.

Results: *This thesis has performed extensive AMG extraction efforts using the document code as a starting-point for extended AMG efforts. This thesis has shown how the document code based approach could gain access to visible and non-visible document content, and how such algorithms can be used in combination with other AMG algorithms. This thesis has documented this by combining usage of the document code approach with traditional (OCR) document section recognition, harvesting and Natural Language processing. This freedom to combine AMG approaches marks a significant step towards enabling automatic generation of metadata from diverse documents.*

RO6: Explore the possibilities for using AMG technologies to assist in generation of advanced and complex to create resources, such as LOs in the SCORM format.

Results: *This thesis has demonstrated automated generation of high-quality SCORM-based learning objects with extensive IEEE LOM metadata descriptions. This has been achieved by performing a series of AMG efforts based around usage of the document code, and by analysing contextual data to give the amount and quality of educational metadata a significant boost. This thesis has presented how the original document and the automatically generated IEEE LOM metadata description can be automatically combined in order to generate a new SCORM LO with high quality metadata.*

5.4 Recommendations

Four main recommendations can be made based on this research:

- **Automatic metadata generation offers a powerful information retrieval tool.** Take advantage of the possibilities that are present for automatically generating metadata based on the documents and document context descriptions. There are extensive opportunities in which to use AMG technologies to generate metadata records for more efficient document retrieval possibilities. Such efforts can be undertaken with documents from both system controlled environments and from stand-alone documents.
- **Be critical of all data sources and AMG methods.** Data sources should be validated before accepting them as use with AMG. Content creation software and their users can create false metadata and intellectual document content that does not match with the document's template and document formatting. The template used to create the document can also contribute with false data. The inconsistencies within the data source need to be documented in order to generate the desired entities while avoiding the generation of false entities. AMG methods based on the document code, visual characteristics, the natural language approach and the harvesting of embedded metadata all have their own strengths and weaknesses. The best way to exploit the possibilities that are present using this approach is by selecting the best candidate AMG method or combination of methods to generate metadata of the highest possible quality.
- **Data sources that are to be used by AMG algorithms should be standardized.** Document templates for use in an organization should also be standardized. The creation of individually identifiable document templates allows the user to create the desired document without having to create a new, user specific document template standard. The commonly available templates should contain the properties desired by users, such as visible presentation and the content in the document. Avoid the inclusion of template content that will become false data once the user stores the template as a new document, e.g. default titles and author names. However, the templates used should be uniquely identifiable so that the most desired AMG efforts can be adapted to that specific template. Users need to know how their efforts are reflected in the resulting documents, and that these data are used as data sources for generating document metadata. Metadata can be used as the basis for re-locating and efficient sharing of knowledge within the organization.

- **Document content should be verified.** If it is possible to verify the document content, this can increase the quality of the document and the quality of the automatically generated metadata. Such verification should be performed when the document is created in order to ensure that the content created by the user is valid and to allow the user to make corrections to presented data. In cases where there are restricted value spaces, only the permitted content should be includable; these restrictions can be enforced by using pull-down lists, click-boxes, and lists of selections and by validating submitted data, such as undertaking a validity check of hyperlinks.

5.5 Future work

There is an extensive amount of research currently being performed regarding AMG and related topics. Still, on day-to-day, real-world basis, the practical benefits of AMG has still a long way off in order to become publically available. This thesis would like to propose the following topics for future work:

- **Research on unstructured, non-homogeneous document collections:** Current research is focused on document collections with visually and subject-wise similar documents. As a result, developed algorithms have limited usability outside of their intended usage area. By basing research on unstructured documents, more general purpose, cross-subject usage areas could be explored.

- **Research on key metadata elements:** Current research is focused on generation of keyword and subject entities. For identification of documents three other elements are commonly highly promoted: The document title, author and creation/modification date. Hardly any research is currently focused on generating such entities from other than highly structured document collections.

- **Research on multi-linguistic documents:** Research on the use of multi-linguistic documents in generating of semantic metadata using natural language approaches: Such as by generating keywords, descriptions and classifications, and by using technologies such as thesauri and ontology on a dataset of multi-linguistic documents.

- **Research on MS Office documents:** A vast number of documents are created in Microsoft Word, Excel and PowerPoint document formats. Still, current AMG-research is commonly conducted on HTML- and occasionally PDF-documents. Research on MS Office documents would extend the usage area of AMG algorithms from being academic tools, to benefit home and business (intranet) retrieval and usage of documents.

- **Promotion of high quality metadata in search engines:** Today's commercial search engines have quality issues regarding the presented entities describing document in MS Office and PDF file formats. Future work should include analysis of the user experienced query results when promoting higher quality metadata though the search engines query results.

- **Automatic generation of Learning Objects:** The amount of publically available Learning Objects on internet is limited and not growing particularly fast (if it actually is growing). This thesis, Meyer et al. [103] and Motelet et al [105] have shown that SCORM-based Learning Objects can be automatically generated based on an existing resource (document) and AMG efforts. There is a need to scale up such efforts, in order to study generation of a larger collection of Learning Objects and the consequences such generation has on the automatically generated metadata.

5.6 Concluding Remarks

Manual creation of metadata is tedious work. It requires extensive knowledge and time from the metadata author. Manual creation of metadata doesn't scale. At the same time our document collections are growing faster than ever. In order to locate the right document, we need metadata. That metadata must be generated, and most likely automatically generated. There are extensive opportunities to continue studies of AMG as the need for AMG will only increase. AMG will be essential for efficient sharing of knowledge in the future.

6. References

[1] ADL, *Alexandria Digital Library Project*, University of California, 2006, http://www.alexandria.ucsb.edu/research/about/history.htm

[2] ADL, *Sharable Content Object Reference Model (SCORM) 2004 3rd Edition Documentation Suite*, 2006, http://www.adlnet.gov/downloads/AuthNotReqd.aspx?FileName=SCORM.2004.3ED.DocSuite.zip&ID=237

[3] ADL, *Advanced Distributed Learning*, 2008, http://www.adlnet.gov

[4] Adobe, *PDF Reference and Related Documentation*, Adobe® Acrobat® 8.1 implementation of the PDF specification, 2007, http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference.pdf

[5] Advanced International Translations, *Word Count and Character Count Software: AnyCount*, 2008, http://www.anycount.com/

[6] ALI, *Apple learning interchange*, 2000, http://ali.apple.com/

[7] H.S. Al-Khalifa, H.C. Davis, H.C., *FAsTA: A Folksonomy-Based Automatic Metadata Generator*, EC-TEL 2007 - Second European Conference on Technology Enhanced Learning, 17-20 September, 2007, Crete, Greece, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.7858

[8] H.S. Al-Khalifa, H.C. Davis, *Replacing the Monolithic LOM: A Folksonomic Approach*, In Proceedings of The 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), 2007, Niigata, Japan, http://eprints.ecs.soton.ac.uk/13882/01/ICALT-07.pdf

[9] J. Allan, *Automatic Hypertext Link Typing*, Hypertext '96, Washington DC USA, ACM 1996, ISBN: 0-89791-778-2/96/03, 1996.

[10] ARIADNE, *ARIADNE Foundation for the European Knowledge Pool*, 2008, http://www.ariadne-eu.org/

[11] D. Bainbridge, D. McKay, I.H. Witten, *Greenstone Digital Library Developer's Guide*, Department of Computer Science, University of Waikato, New Zealand, Greenstone gsdl-2.50, March 2004. http://www.greenstone.org/developers-guide

[12] S. Bateman, C. Brooks, G. McCalla, P. Brusilovsky, *Applying Collaborative Tagging to E-Learning*, In Proceedings of the 16th International World Wide Web Conference (WWW2007), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.8892&rep=rep1&type=pdf

[13] T. Berners-Lee, *Metadata Architecture*, 1997, http://www.w3.org/DesignIssues/Metadata.html

[14] K. Bird, the Jorum Team, *Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems*, 31st March 2006, Version 1.0, Final, Signed off by JISC and Intrallect July 2006, http://www.jorum.ac.uk/docs/pdf/automated_metadata_report.pdf

[15] Blackboard, *Blackboard » Educate. Innovate. Everywhere*, 2008, http://www.blackboard.com

[16] B. Boguraev, M. Neff, *Lexical Cohesion, Discourse Segmentation and Document Summarization*, RIAO, 2000, http://citeseer.ist.psu.edu/cache/papers/cs/20369/http:zSzzSz133.23.229.11zSz~ysuzuki zSzProceedingsallzSzRIAO2000zSzThursdayzSz79DO3.pdf/lexical-cohesion-discourse-segmentation.pdf

[17] T.R. Bruce, D.I. Hillmann, *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*", ALA Editions, In Metadata in Practice, D. Hillmann & E Westbrooks, eds., 2004, URI: http://hdl.handle.net/1813/7895, ISSN: 0-8389-0882-9, http://ecommons.library.cornell.edu/bitstream/1813/7895/1/Bruce_Hillmann_corr_final.doc

[18] J.D. Byrum, O.M.A. Madison, *Reflections on the Goals, Concepts and Recommendations of the IFLA Study on Functional Requirements of Bibliographic Records*, FRBR (Functional requirements for bibliographic records) SEMINAR - Florence, 27-28 January 2000, Associazione Italiana Biblioteche, http://www.aib.it/aib/sezioni/toscana/conf/frbr/byrmadis.htm

[19] CanCore, *Guidelines*, 2004, http://www.cancore.ca/en/guidelines.html

[20] K. Cardinaels, E. Duval, H. Olivié, *A Formal Model for Learning Object Metadata*, EC-TEL 2006, Lecture Notes in Computer Science, Volume 4227/2006, pp. 74-87, ISSN 0302-9743 (Print) 1611-3349 (Online), ISBN: 978-3-540-45777-0, Springer Berlin / Heidelberg, DOI: 10.1007/11876663, http://www.springerlink.com/content/m63808qn53241r24/

[21] K. Cardinaels, M. Meire, E. Duval, *Automating metadata generation: the simple indexing interface*, Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, pp.548-556, 2005, ISBN:1-59593-046-9, http://portal.acm.org/ft_gateway.cfm?id=1060825&type=pdf&coll=GUIDE&dl=GUIDE&CFID=37384861&CFTOKEN=19135172

[22] CETIS, *UK LOM Core v 0.2*, 2004, http://www.cetis.ac.uk/profiles/uklomcore/uklomcore_v0p2_may04.doc

[23] A. Crystal, P. Land, *Metadata and Search*, Report on the Global Corporate Circle DCMI 2003 Workshop, held at the 2003 Dublin Core Conference: Supporting

Communities of Discourse and Practice—Metadata Research and Applications. Seattle, Washington: September 28 - October 3, 2003, http://dublincore.org/groups/corporate/Seattle/

[24] DCMI, *Dublin Core Qualifiers*, 2000, http://www.dublincore.org/documents/2000/07/11/dcmes-qualifiers/

[25] DCMI, *DC-Education Application Profile*, 2006, http://projects.ischool.washington.edu/sasutton/dcmi/DC-EdAP-7-18-06.html

[26] DCMI, *DCMI Education Community*, 2006, http://dublincore.org/groups/education/

[27] DCMI, *Dublin Core Metadata Element Set, Version 1.1*, 2006, http://dublincore.org/documents/dces/

[28] DCMI, *Dublin Core Metadata Initiative (DCMI)*, 2008, http://dublincore.org/

[29] DCMI Usage Board, *DCMI Type Vocabulary*, 2006, http://dublincore.org/documents/dcmi-type-vocabulary/

[30] F. de Jong, D. Giuliani, M. Kaiser, S. Kopf, L. Lamel, J. Oomen, F. Rauh, M. Rendina, P. Savino, S. Schneider, H. Wactlar, *ECHO – European Chronicles on Line (IST-1999-11994), final report, March 2003*, Editor Savino, P., Information Society Technologies, 2003, http://pc-erato2.iei.pi.cnr.it/echo/documents/public/Final%20Report.pdf

[31] E. Di Nitto, L. Mainetti, M. Monga, L. Sbattella, R. Tedesco, *Supporting Interoperability and Reusability of Learning Objects: The Virtual Campus Approach*, Educational Technology & Society, 9 (2), 33-50, 2006, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.1766 .

[32] DLESE, *ADN framework*, Digital Library for Earth System Education (DLESE), 16.12.2004, http://www.dlese.org/Metadata/adn-item/history.htm

[33] DLESE, *Digital Library for Earth System Education*, NSDL, 2008, http://www.dlese.org

[34] DLI, *Digital Library Initiative Phase 2*, NSF, 2003, http://www.dli2.nsf.gov/

[35] M. Dobreva, Y. Kim, *Automatic Metadata Generation – Use Cases: File Format Metadata (Definitive for Preservation)*, the JISC MOSAIC project, 2009, http://www.intrallect.com/index.php/intrallect/content/download/947/3986/file/File_Format_Metadata.pdf

[36] E. Duval, W. Hodgins, *Making metadata go away: Hiding everything but the benefits*, Keynote address at DC-2004, Shanghai, China, 2004, http://students.washington.edu/jtennis/dcconf/Paper_15.pdf

[37] E. Duval, D. Massart, F. Van Assche, S. Ternier, S. Hartinger, *Content enrichment toolbox*, Production version, Project Number: ECP 2005 EDU 038103, Project Title: MELT, Deliverable Number: D 2.2, Version: 1.0, Deliverable Type*: PP, Nature of the Deliverable**: P, Contractual Date of Delivery: 31 March 2007, Actual Date of Delivery: 31 March 2007, Work-Package contributing to the Deliverable: WP 2, 2007, http://eunbrux09.eun.org/shared/data/melt/MELT_D2P2_final.pdf

[38] DWEL, *Digital Water Education Library – About the DWEL collection*, The Center for Science, Mathematics, and Technology Education, Colorado State University, 2003, http://www.csmate.colostate.edu/dwel/

[39] ECHO, *European Union ECHO European CHronicles On-line*, 2005, http://pc-erato2.iei.pi.cnr.it/echo/

[40] L.F.H. Edvardsen, I.T. Sølvberg, *Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment*, Proc.of WEBIST 2007, March 3-6, ISBN 978-972-8865-77-1, pp. 427-432, Springer, 2007

[41] L.F.H. Edvardsen, I.T. Sølvberg, *Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing?*, Proc. of IEEE DEST, IEEE Computer Society 2010 ISBN 978-1-4244-5553-9. pp. 216-221, 2010

[42] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg, H. Trætteberg, *Using the structural content of documents to automatically generate quality metadata*, Proc. of Webist 2009, March 23-26, pp. 354-363, ISBN: 978-989-8111-83-8, ACM, 2009

[43] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg, H. Trætteberg, *Automatically generating high quality metadata by analyzing the document code of common file types*, Proc. of JCDL 2009, June 15-19, ACM, 2009

[44] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg, H. Trætteberg, *Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects*. Proc. of IEEE DEST 2009, June 1-3, INSPEC, 2009, ISBN: 978-1-4244-2345-3, 10.1109/DEST.2009.5276729

[45] ESCOT, *Educational software components of tomorrow*, 2000, http://www.escot.org/

[46] eStandard, *Norsk LOM-profil – NORLOM. Versjon 1.0*, 2005, http://www.estandard.no/norlom/v1.0/NORLOM_v1_0_mars_2005.pdf

[47] EUN, *European Partners*, 2007, http://dotsafe.eun.org/dotsafe.eun.org/eun.org2/eun/en/About_eschoolnet/sub_area89d6.html?sa=757

[48] FIRE/LRE, *The EUN Learning Document Exchange Metadata Application Profile Version 3.0 June 2007*, 2007, http://fire.eun.org/LRE-AP-3.0.pdf

[49] P. Flynn, L. Zhou, K. Maly, S. Zeil, M. Zubair, *Automated Template-Based Metadata Extraction Architecture*, ICADL 2007

[50] P. Foster, M. Kraner, A. Graziano, S.P. Romano, *GESTALT - Project AC367. Getting Educational Systems Talking Across Leading-Edge Technologies. Work Package 4. D0401 Courseware Metadata Design V3 (GEMSTONES)*, British Telecommunications plc 2000, 2000, http://www.fdgroup.co.uk/gestalt/D0401_3.pdf

[51] FREPA, *frepa.blog - On e-learning and Learning Technology » Blog Archive » SWE-LOM: a Swedish LOM application profile*, 2006, http://www.frepa.org/wp/2006/06/28/swe-lom-a-swdish-lom-application-profile/

[52] N. Friesen, *Final Report on the "International LOM Survey"*, CAC JTC1/SC36 Document No: 36C087 2004-08-25, 2004, http://jtc1sc36.org/doc/36N0871.pdf

[53] GEM, *Listing of GEM 2.0 top-level elements (i.e., elements that are not refinements of other elements)*, version June 1, 2004, http://www.thegateway.org/about/documentation/metadataElements

[54] GESTALT, *GESTALT*, Fretwell Downing Education, 1999, http://www.fdgroup.co.uk/gestalt

[55] A.J. Gilliland-Swetland, *Setting the Stage*, 1998, http://www.mokk.bme.hu/mediatervezo/targyak/metainfo/2004/swetland.pdf

[56] G. Giuffrida, E.C. Shek, J. Yang, *Knowledge-Based Metadata Extraction from PostScript Files*, Digital Libraries, San Antonio, Tx, ACM 1-581 13-231-X/00/0006, 2000, http://portal.acm.org/ft_gateway.cfm?id=336639&type=pdf&coll=GUIDE&dl=GUIDE&CFID=37040527&CFTOKEN=25109993

[57] Google, *Google*, 2008, http://www.google.com

[58] Google, *Google*, 2011, http://www.google.com

[59] Google Desktop, *Google Desktop Download*, 2008, http://desktop.google.com

[60] J. Greenberg, K. Spurgin, A. Crystal, M. Cronquist, A. Wilson, *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project*, UNC School of information and library science, 2005, http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

[61] J. Greenberg, *Metadata and the World Wide Web*, M.S. Drake (Ed.) Encyclopedia of library and information science (2nd ed.) (pp.1876-1888). New York: Marcel Dekker, Inc., 2003, http://www.ils.unc.edu/mrc/pdf/greenberg03metadata.pdf

[62] J. Greenberg, *Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications*, Journal of Internet Cataloging, 6(4): 59-82., 2004, http://www.ils.unc.edu/mrc/pdf/greenberg04metadata.pdf

[63] Greenstone, *Source only distribution*, 2007, http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected)

[64] Greenstone, *Greenstone Digital Library software*, 2008, http://www.greenstone.org

[65] M. Hämäläinen, B.A. Whinston, S. Vishik, S., *Electronic markets for learning: education brokerages on the Internet*, Communications of the ACM archive, Volume 39, Issue 6 (June 1996), pp. 51 – 58, ISSN:0001-0782, 1996, http://portal.acm.org/ft_gateway.cfm?id=228513&type=pdf&coll=portal&dl=ACM&CFID=33796525&CFTOKEN=11974524

[66] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E.A. Fox, *Automatic Document Metadata Extraction using Support Vector Machines*, Proc. of the 2003 Joint Conference on Digital Libraries (JCDL'03), 2003, ISBN: 0-7695-1939-3/03

[67] P.B. Heidorn, Q. Wei, *Automatic Metadata Extraction from Museum Specimen Labels*, Proc. Int'l Conf. on Dublin Core and Metadata Applications 2008, 2008.

[68] Hillbilly Software, *Hillbilly Software – PDF Reverser*, 2008, http://www.hillbillysoft.com/pdfreverser.html

[69] D. Hillmann, *Using Dublin Core*, Dublin Core Metadata Initiative, 2001, http://dublincore.org/documents/2001/04/12/usageguide

[70] M. Hlava, *Breaking Down Automatic Metadata Generation/Extraction*, 2011, http://taxodiary.com/2011/06/breaking-down-automatic-metadata-generationextraction-2/

[71] Y. Hu, H. Li, Y. Cao, L. Teng, D. Meyerzon, Q. Zheng, *Automatic extraction of titles from general documents using machine learning*, Information Processing and Management: an International Journal, Volume 42, Issue 5 (September 2006), pp.1276-1293, 2006, ISSN:0306-4573, http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC8-4J5T8BP-1&_user=586462&_coverDate=09%2F30%2F2006&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000030078&_version=1&_urlVersion=0&_userid=586462&md5=84ab0ed9a1ccb6b8c1bd851ed5833c3d

[72] IEEE LTSC, *LOM working draft v4.1*, 2000, http://ltsc.ieee.org/doc/wg12/LOMv4.1.htm

[73] IEEE LTSC, *IEEE Standard for Learning Object Metadata. IEEE Standard 1484.12.1*, Institute of Electrical and Electronics Engineers, New York, 2002. http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

[74] IEEE LTSC, *IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for*

Learning Object Metadata, WG12: Related Materials, 2005, http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

[75] IEEE LTSC, *LTSC Home Page – IEEE Learning Technology Standards Committee*, 2008, http://ieeeltsc.org/

[76] IFLA, *Functional Requirements for Bibliographic Records – Final Report*, UBCIM Publications - New Series Vol 19, IFLA Study Group on the Functional Requirements for Bibliographic Records, K. G. Saur München 1998, UBCIM Publications, New Series, 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[77] IFLA, *International Federation of Library Associations and Institutions*, 2008, http://www.ifla.org/

[78] IMS, *IMS Learning Document Meta-data Best Practices and Implementation Guide Version 1.0 - Final Specification*, 20 August 1999, IMS Global Learning Consortium, Inc., 1999, http://www.imsglobal.org/metadata/mdbest01.html

[79] IMS, *Welcome to IMS Global Learning Consortium, Inc.*, 2008, http://www.imsglobal.org/

[80] Internet Mail Consortium, *vCard and vCalendar*, 2008, http://www.imc.org/pdi/

[81] It's learning, *It's learning*, 2008, http://www.itslearning.com

[82] C. Jenkins, D. Inman, *Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF*, 0-7695-1022-1/01, 2001 IEEE, 2001, http://ieeexplore.ieee.org/iel5/7499/20393/00942183.pdf?arnumber=942183

[83] C. Jenkins, M. Jackson, P. Burden, J. Wallis, *Automatic RDF Metadata Generation for Document Discovery*, The Eight International World Wide Web Conference, 1999, http://www8.org/w8-papers/2c-search-discover/automatic/automatic.html

[84] Jorum, *Jorum Home*, 2008, http://www.jorum.ac.uk

[85] A. Kawtrakul, C. Yingsaeree, *A Unified Framework for Automatic Metadata Extraction from Electronic Document*, Proceedings of IADLC2005 (The International Advanced Digital Library Conference) (25-26 August 2005), pp. 71-77. 2005, http://iadlc.nul.nagoya-u.ac.jp/archives/IADLC2005/kawrtrakul.pdf

[86] Y. Kim, S. Ross, *Automating Metadata Extraction: Genre Classification Poster*, UK e-Science All Hands Meeting, 2006, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6930

[87] V. Kløvstad, T. Kristiansen, *Skolens Digitale Tilstand 2003*, "The schools digital state", ITU Monitor, Rapport 1 2004, Forsknings- og kompetansenettverk for IT i utdanning (ITU). The University of Oslo. 2004, ISBN 82-7947-024-7. ISSN 1503-8432. Also available at http://www.itu.no/Filer/fil_ITU_Monitor_rapport1.pdf

[88] L. Kolås, L.F.H. Edvardsen, L.M. Hokstad, *Bruk av It's learning ved NTNU – en kvantitativ og kvalitativ studie*", NTNU, 2008, http://www.ntnu.no/c/document_library/get_file?uuid=dbd0c675-7b3a-4ca6-b602-242d4067b250&groupId=524136

[89] D.A. Koutsomitropoulos, A.D. Alexopoulos, G.D. Solomou, T.S. Papatheodorou, *The Use of Metadata for Educational Resources in Digital Repositories: Practices and Perspectives*, D-Lib Magazine, January/February 2010, Volume 16, Number 1/2, 2010, http://www.dlib.org/dlib/january10/kout/01kout.html

[90] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage Publications Inc., 2004, ISBN 0761915451, http://books.google.com/books?hl=en&lr=&id=q657o3M3C8cC&oi=fnd&pg=PR13&dq=character+word+page+understanding+amount+of+content&ots=bIhbw1L7xY&sig=ZEbOvADd8a_ae3b4IAolXi0dJIU

[91] H. Li, Y. Cao, J. Xu, Y. Hu, S. Li, D. Meyerzon, *A new approach to intranet search based on information extraction*, Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, Pages: 460 – 468, 2005, ISBN:1-59593-140-6, ACM New York, NY, USA, http://portal.acm.org/ft_gateway.cfm?id=1099685&type=pdf&coll=GUIDE&dl=GUIDE&CFID=43046171&CFTOKEN=22468096

[92] Y. Li, C. Dorai, R. Farrell, *Creating MAGIC: system for generating learning object metadata for instructional content*, Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, pp.367-370, 2005, ISBN:1-59593-044-2, http://portal.acm.org/ft_gateway.cfm?id=1101227&type=pdf&coll=GUIDE&dl=GUIDE&CFID=1299051&CFTOKEN=15183150

[93] Y. Li, Q. Zhu, Y. Cao, *Automatic Metadata Generation based on Neural Network*, Conference'04, Month 1-2, 2004, Pudong, Shanghai, China, 2004 ACM, ISBN: 1-58113-955-1, http://portal.acm.org/ft_gateway.cfm?id=1046330&type=pdf&coll=GUIDE&dl=GUIDE&CFID=37041914&CFTOKEN=82143250

[94] E.D. Liddy, E. Allen, S. Harwell, S. Corieri, O. Yilmazel, N.E. Ozgencil, A. Diekema, N.J. McCracken, J. Silverstein, S.A. Sutton, *Automatic metadata generation and evaluation*, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11–15, Tampere, Finland, ACM Press, New York, pp.401–402. 2002, http://portal.acm.org/ft_gateway.cfm?id=564464&type=pdf&coll=GUIDE&dl=GUIDE&CFID=37165125&CFTOKEN=53575886

[95] C.A. Linderoth, A. Bandholm, B. Christensen-Dalsgaard, G. Berger, *The European Schoolnet: An Attempt to Share Information and Services*, Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes In Computer Science; Vol. 1513, pp. 683-684, 1998, ISBN:3-540-65101-2, Springer-Verlag

[96] O.I. Lindland, G. Sindre, A. Sølvberg, *Understanding Quality in Conceptual Modeling*, IEEE Software, march 1994, Volume: 11, Issue: 2, pp. 42-49, 1994, ISSN: 0740-7459, DOI: 10.1109/52.268955, http://ieeexplore.ieee.org/iel1/52/6703/00268955.pdf?tp=&isnumber=&arnumber=268955

[97] Y. Liu, K. Bai, P. Mitra, C.L. Giles, *TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries*, JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada, ACM 978-1-59593-644-8/07/0006, 2007, http://portal.acm.org/ft_gateway.cfm?id=1255193&type=pdf&coll=GUIDE&dl=GUIDE&CFID=37132119&CFTOKEN=72486418

[98] LOMGen, *LOMGen*, 2006, http://www.cs.unb.ca/agentmatcher/LOMGen.html

[99] M. Meire, X. Ochoa, E. Duval, *SAmgI: Automatic Metadata Generation v2.0*, In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007 (pp. 1195-1204). Chesapeake, VA: AACE, 2007, http://www.editlib.org/index.cfm/files/paper_25528.pdf?fuseaction=Reader.DownloadFullText&paper_id=25528

[100] MELT, *MELT – Learning Documents for schools*, 2007, http://info.melt-project.eu/ww/en/pub/melt_project/welcome.htm

[101] MERLOT, *Multimedia educational document for learning and on-line teaching*, 2008, http://www.merlot.org/

[102] M.D. Merrill, Z. Li, M. Jones, *Instructional transaction theory: An introduction. Educational Technology*, Published Educational Technology, 1991, 31(6), 7-12, http://cito.byuh.edu/merrill/text/papers/ITT_Intro.PDF

[103] M. Meyer, C. Rensing, R. Steinmetz, *Categorizing Learning Objects Based On Wikipedia as Substitute Corpus*. Proc. of LODE-07, the 1 st Int. Workshop on Learning Object Discovery & Exchange, 2007, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.109.7498

[104] E. Miller, *An Introduction to the Resource Description Framework*, D-Lib Magazine, May 1998, 1998, ISSN 1082-9873, http://www.dlib.org/dlib/may98/miller/05miller.html

[105] O. Motelet, N. Baloian, *Hybrid System for Generating Learning Object Metadata*, Proc. of ICALT-06, 2006, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.1449

[106] M. Naaman, R.B. Yeh, H. Garcia-Molina, A. Peapcke, *Leveraging context to resolve identity in photo albums*. In Proceedings of the 5th Joint Conference on Digital Libraries (JCDL'05), 2005

[107] NASA, *science@nasa*, 2008, http://science.hq.nasa.gov/earth-sun/index.html

[108] F. Neven, E. Duval, *Reusable learning objects: a survey of LOM-based repositories*, Proceedings of the tenth ACM international conference on Multimedia, Juan-les-Pins, France, pp. 291 – 294, 2002, ISBN:1-58113-620-X, ACM Press New York, NY, USA, http://portal.acm.org/ft_gateway.cfm?id=641067&type=pdf&coll=GUIDE&dl=GUIDE&CFID=29374048&CFTOKEN=94602360

[109] M. Ni, *Automatic Extraction of Author Self Contributed Metadata for Electronic Thesis and Dissertations*, Master's paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina, May 2004, 2004, http://etd.ils.unc.edu:8080/dspace/bitstream/1901/88/1/MaoNi.pdf

[110] M. Nilsson, M. Palmér, A. Naeve, *Semantic Web Metadata for e-Learning - Some Architectural Guidelines*, 2002, http://www2002.org/CDROM/alternate/744/index.html

[111] NISO, *Understanding Metadata*, NISO (National Information Standards Organization), 2004, ISBN: 1-880124-62-9, http://www.niso.org/standards/documents/UnderstandingMetadata.pdf

[112] NSDL, *Quick Table of NSDL Metadata Elements*, 2001, http://standards.comm.nsdlib.org/ElementsTable2.html

[113] NSF, *National Science Foundation*, 2008, http://www.nsf.gov/

[114] X. Ochoa, K. Cardinaels, M. Meire, E. Duval, *Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories*, Proceedings of EDMedia 2005, Montreal, Canada, pp. 1407-1414, 2005, http://ariadne.cs.kuleuven.ac.be/amg/publicationsFiles/paperAMG2.doc

[115] OCLC, *Dewey services – Dewey Decimal Classification*, 2008, http://www.oclc.org/dewey/

[116] Open Archives Initiative, *Protocol for Metadata Harvesting – v.2.0*, 2004, http://www.openarchives.org/OAI/openarchivesprotocol.html

[117] J. Peneva, G. Totkov, P. Stanchev, E. Shoikova, *The Metaspeed project – Progress report*, The 7 Annual International Conference on Computer Science and Education in Computer Science (CSECS 2011), July 06-10 2011, Sofia, Bulgaria, 2011, http://eprints.nbu.bg/801/1/CSECS2011-Peneva_et-al..pdf

[118] M.M. Recker, D.A. Wiley, *A non-authoritative educational metadata ontology for filtering and recommending learning objects*, Journal of Interactive Learning Environments, Swets and Zeitlinger, The Netherlands, 2001, http://www.informaworld.com/smpp/ftinterface?content=a725291219&rt=0&format=pdf

[119] C.M. Reigeluth, L.M. Nelson, *A new paradigm of ISD?*, R. C. Branch & B. B. Minor (Eds.), Educational media and technology yearbook (Vol. 22, pp. 24-35). Englewood, CO: Libraries Unlimited, 1997

[120] M.A. Rodriguez, J. Bollen, H. Van de Sompel, *Automatic Metadata Generation Using Associative Networks*, ACM Transactions on Information Systems, Vol. 27, No. 2, Article 7, 2009, http://public.lanl.gov/herbertv/papers/Papers/2009/TOISrodriguez.pdf

[121] Scirus, *Scirus – for scientific information only*, 2011, http://www.scirus.com

[122] K. Seymore, A. McCallum, R. Rosenfeld, *Learning hidden Markov model structure for information extraction*. Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction, pages 37-42, 1999, http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/kseymore/papers/ie_aaai99.ps.gz

[123] A. Singh, H. Boley, V.C. Bhavsar, *LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology*, National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004, http://www.cs.unb.ca/agentmatcher/LOMGen/paper/LOMGen-29Mar2004.ppt

[124] Skolenettet, *Skolenettet*, Utdanningsdirektoratet, 2008, http://www.skolenettet.no

[125] Smart PC Solutions, *Smart PC Solutions*, 2008, http://www.smartpctools.com/en/index.html

[126] Soft Experience, *Metadata Miner Catalogue PRO software*, 2008, http://peccatte.karefil.com/software/Catalogue/MetadataMiner.htm

[127] Springer, *Information for LNCS Authors*, 2007, http://www.springer.com/east/home/computer/lncs?SGWID=5-164-7-72376-0&teaserId=45515&CENTER_ID=73062

[128] S. Syn, M. Spring, *Can a System Make Novice Users Experts? Important Factors for Automatic Metadata Generation Systems*, Proceedings of the international conference on Dublin Core and Metadata applications, pp.140-150, 2007, http://www.dcmipubs.org/ojs/index.php/pubs/article/view/32

[129] I.T. Sølvberg, L.F.H. Edvardsen, *Creating Metadata is a Costly Manual Process – And it can be Automated,* In: Antony Jose (ed.) "Digital Libraries and Knowledge Organizations." Macmillan Publishers India Ltd., 2012. 2011, ISBN 978-935-059-076-8. Pp. 356-362.

[130] A. Takasu, *Bibliographic attribute extraction from erroneous references based on a statistical model*. In Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 49 – 60). New York: ACM Press, 2003

[131] The Open Group, *Universal Unique Identifier*, 1997, http://www.opengroup.org/onlinepubs/9629399/apdxa.htm

[132] J. Voss, *Tagging, Folksonomy & Co - Renaissance of Manual Indexing?*. Proceedings of the International Symposium of Information Science, Cologne, Germany. pp.234–254, 2007, http://arxiv.org/pdf/cs/0701072

[133] D.A. Wiley, *Learning Object design and sequencing theory*, A dissertation submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Instructional Psychology and Technology, Brigham Young University, June 2000, http://wiley.ed.usu.edu/docs/dissertation.pdf

[134] Y. Xue, Y. Hu, G. Xin, R. Song, S. Shi, Y. Cao, C-Y. Lin, H. Li, *Web page title extraction and its application*, Information Processing & Management, Volume 43, Issue 5, September 2007, Pages 1332-1347, 2007, doi:10.1016/j.ipm.2006.11.007, http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6VC8-4MVDVSF-2-1&_cdi=5948&_user=586462&_orig=search&_coverDate=09%2F30%2F2007&_sk=999569994&view=c&wchp=dGLbVzb-zSkzS&md5=209fd927cb4c2d20b89c8f1f0f8bd780&ie=/sdarticle.pdf

[135] Yahoo, *Yahoo!*, 2011, http://www.yahoo.com

Appendix A: Papers

P1 Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg, *"Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment"*.

P2 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Automatically generating high quality metadata by analyzing the document code of common file types"*.

P3 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Using the structural content of documents to automatically generate quality metadata"*.

P4 Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg, *"Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing?"*.

P5 Lars Fredrik Høimyr Edvardsen, Ingeborg Torvik Sølvberg, Trond Aalberg and Hallvard Trætteberg, *"Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects"*.

P6 Ingeborg Torvik Sølvberg and Lars Fredrik Høimyr Edvardsen, *"Creating Metadata is a Costly Manual Process – And it can be Automated"*.

P1: Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment

Published in the Proceedings of WEBIST 2007,
March 3-6, 2007, ISBN 978-972-8865-77-1, pp. 427-432, Springer

Lars Fredrik Høimyr Edvardsen and Ingeborg Torvik Sølvberg

Dept. of Computer and Information System,
Norwegian University of Science &Technology,
Sem Sælands vei 7-9, NO-7491, Trondheim, Norway
{lars.edvardsen,ingeborg} @idi.ntnu.no

Keywords: IEEE LOM, Learning Object Metadata, LOM, Learning Object, LO, Learning Management System, LMS, metadata mapping, crosswalk, metadata challenges

Abstract: The world of closed Learning Management Systems (LMS) is being replaced by open systems for sharing and reusing digital Learning Objects (LOs) between users, courses, institutions and countries. This poses new challenges in describing these LOs with detailed and correct metadata. This information background is needed for querying services to perform accurate queries for LO retrieval. In this paper we present metadata specific challenges when converting from a local LMS with proprietary metadata schema to a global metadata schema. We have uncovered extensive LO description possibilities based on the existing, local LMS, registered metadata, its LO types and the local context. Files can contain extensive metadata descriptions, though require special attention. We have confirmed that technologies developed as crosswalks are valid for usage in this projects for a one-time metadata transferral. However, transferring of all local metadata elements can result in incompatibility issues with other LMSs. This, even when keeping with the global metadata schema.

1 INTRODUCTION

The use of digital Learning Objects (LOs) such as slides, figures, exercises and exams are increasing on all educational levels. This is happening all over the world, in use by both students and teachers. The current generation of Learning Management Systems (LMSs) have had limited, if any LO and Learning Object Metadata (LOM) sharing possibilities. A new generation of LMSs is now emerging which allow sharing and reuse of LOs. Their LOM descriptions are the vital information background needed for

querying services to perform accurate queries for LOs. For LMSs this transformation process means converting from a proprietary, local metadata schema to a global schema.

Between intentionally compatible metadata schemas, metadata exchange can be performed lossless. E.g. the national schemas UK LOM Core (Cetis, 2004) (UK) and NORLOM (eStandard, 2005) (Norwegian) are compatibility with the global IEEE LOM (IEEE LTSC, 2005).

For schemas without a pre-intended compatibility, metadata exchange can be more challenging. This is the case for most LMSs. A potential solution is using crosswalks (Chan & Zeng, 2006). Crosswalks are a set of determined equal elements between two schemas. This allow transfer of metadata back and forth between two schema standards, e.g. between Dublin Core and MARC (Library of Congress, 2001). In our work, crosswalks will be used as a one-way tool to transfer existing metadata to the new schema. However, since these schemas are not equal, many-to-one element mappings and many-to-none element mappings can occur. Here the fine-grain metadata schema architecture and existing metadata can get lost when converting. Cases with unequal elements resulting in one-to-many elements need to be addressed.

Metadata mapping is actually an everyday event when converting file formats. Though, it is often hidden from user sight, like when converting MS PowerPoint slides into Adobe PDF print-outs. How the original metadata elements are converted, updated, excluded or replaced by other metadata is determined by the converting software, such as Adobe Distiller.

If files are to be used as a metadata source, this poses special challenges: There are a range of different file formats in use; many have a proprietary metadata schema. Our studies have uncovered extensive differences in how elements are used. As a result, files need to be given special attention if used as a metadata source.

These are all challenges facing the Norwegian University of Science and Technology (NTNU). Here the Local LMS (LLMS) metadata schema will be converted to NORLOM. The LLMS has a proprietary schema with little resemblance to the destination schema. It uses other element names, which can make discovering of existing metadata sources more challenging. It has extensive use of elements not covered by the IEEE LOM. And it has single elements covering multiple IEEE LOM elements. This results in one-to-many, many-to-one and many-to-none element situations. In addition files are a frequently used LO type, resulting in additional metadata challenges when included as a metadata source.

2 The IEEE LOM schema

The IEEE LOM schema is specially adapted to describe LOs. It divides metadata elements into predefined categories: General, Life Cycle, Meta-Metadata, Technical, Educational, Rights, Relation and Annotation. For other metadata, a 9th category Classification can be used. The initial 8 categories open for LO descriptions containing more than 60 different elements, most of them reusable for multiple registrations. This

vastness in numbers and the preciseness of each element poses challenges when moving from a local to this global metadata schema.

The Classification category where created to support a local LO identification schema. It allows creation of local elements within an existing schema structure. Other metadata elements can be included in this category. They are not globally valid, because they only follow a local schema. Re-usage of these metadata can only be performed by the local LMS and other LMSs and services compatible with the local schema.

3 Using an existing LMS as metadata source

3.1 Discovering potential metadata sources within the LLMS

The LLMS is divided into course-specific sections. Each course has a course-profile with information including: course-name, id, year and semester. The course id includes information about the "course owner", such as the university department. Each course has predefined users which must log-in to gain course and LO access. Each user has a profile which includes user name, login-information and e-mail address.

The LLMS has functions for distributing course information. Common usage includes sharing of curriculum lists, slides from lectures, presentations of student assignments, e-mail and chat. The legal types of LOs are note, link, exercise, online test, question (chat) session, report and upload file. Each LO type have specific, predefined properties. All the LO types have administrative metadata: publisher name (creator), folder name, date and title.

The LLMS do not control or check uploaded files. Users can upload any file and store it in a course specific section. The most commonly used file formats are MS Office-based, Adobe PDF and JPEG images. These file types have extensive, custom metadata schemas. This is also true for many other used file formats. Hence files can be an uncertain and complicated metadata source.

3.2 Schema mapping

The LLMS has potentially multiple metadata sources: User-, Course-, Institution- and University profiles, and LOs created within the LLMS, as well as uploaded files.

The metadata elements of these sources should now be transferred to the new, global schema. (Zeng & Xiao, 2001) describes 4 relation types: one-to-one, one-to-many, one-to-none and many-to-one.

One-to-one relations are lossless and are used in crosswalks. Here equivalent element types are mapped as they were the same element type. This includes converting between equal schemas with different formatting, e.g. between date formatting: year, month, day vs. day, month and year.

One-to-many elements indicate that the destination schema has finer grain allowing more precise metadata descriptions. Common elements include descriptions of local custom elements.

One-to-none elements indicate a direct loss of metadata from the existing schema. Within any converting process, an aim would be to avoid losing data. Effort should hence be enforced to avoid this issue.

Many-to-one elements indicate a less grained destination schema. This can result in less detailed metadata descriptions.

3.3 One-to-one elements

The precise definition of the LLMS' LO types, except files, can be used to create crosswalks or one-to-one element relations. This is because of equality between some of the predefined LLMS metadata schema elements and the defined targeting schema elements. Between the two schemas there are equal elements, like shown in Table 1.

Table 1: Title

LLMS metadata	LLMS title = Exercise nr 2
IEEE LOM metadata	1.2 Title = Exercise nr 2

3.4 One-to-many elements

Within the LLMS there is extensive use of local information which is not explicitly described. Moving from a local LMS schema to a global schema will require describing the local schema and its surroundings in the global schema's terms. This includes course specific elements and interpretation of local course characteristics. These can be collected in a course profile allowing LOs created or uploaded to the course to take advantage of the course profile. Candidate course profile elements include course description and its primary user group, as shown in Table 2.

Table 2: Course context

LLMS metadata	LLMS course context = IT3805
IEEE LOM metadata	5.5 Intended End User Role = Learner 5.8 Difficulty = Very difficult 5.11 Language = NO 9.2.2 Taxon = {[“Institute”, “IDI”]} 9.2.2 Taxon = {[“Course”, “IT3805”]}

Other candidate elements can be set at a general level for the University as a whole, at Institute and department levels, down to low level, fine grained elements set by individual course lecturers. These profiles can describe practical usage properties of the LMS and all its users, schema name, policy and other politically tuned elements. See Table 3 for an example.

Table 3: University context

LLMS metadata	LLMS University context = NTNU
IEEE LOM metadata	5.6 Context = Higher education 5.7 Typical age range = 18- 9.2.2 Taxon = {[“University”, “NTNU”]}

Some local elements require usage of multiple global elements to cover the local description. E.g. the LLMS’ “Exercise” LO has a range of properties not covered by an individual LO type in IEEE LOM. To fully describe the “Exercise” LO multiple IEEE LOM elements have to be created, as shown in Table 4.

Table 4: LO type description

LLMS metadata	LLMS LO type = Exercise
IEEE LOM metadata	4.1 Format = text/html 5.1 Interactivity type = Active 5.2 Learning Resource type = Exercise 5.3 Interactivity level = High

3.5 One-to-none elements

The issue of one-to-none elements poses a danger of losing data when converting from a local to a global schema. One example is when converting the "Exercise" LO type. It has specific elements specifying if an exercise is mandatory and its delivery date, see Table 5. Such elements are not covered by the IEEE LOM schema.

Table 5: Local elements

LLMS metadata	LO: Obligatory = Yes LO: Final delivery date = 01.10.2006
IEEE LOM metadata	-

For these two exemplified elements and other elements without an equivalent IEEE LOM element, there are two lossless possibilities: Use of an unstructured general description or extend the IEEE LOM schema with custom elements. The first solution results in a many-to-one element situation with loss of precision within the schema as a result. Table 6 shows this scenario by storing the existing element names and entities as a merged text string within the General Description element.

Table 6: Using 1.4 Description for local elements

LLMS metadata	LO: Obligatory = Yes LO: Final delivery date = 01.10.2006
IEEE LOM metadata	1.4 Description = "Obligatory = Yes" 1.4 Description = "Final delivery date = 01.10.2006"

An alternative can be to use the Classification category to extend the global schema. This can result in a lossless schema and metadata coverage, see Table 7 ("NO" referring to language, other string elements refer to element content).

Table 7: Using Classification for local elements

LLMS metadata	LO: Obligatory = Yes
IEEE LOM metadata	9.1 Purpose = Educational Objective 9.2.1 Source = (”NO”,”NTNU LMS”) 9.2.2 Taxon = {[”Obligatory”, ”YES”]}

Use of the Classification category can resolve the missing global elements issue by creating local elements. Simultaneously it looses schema compatibility with other LMSs for these specific elements. One of the intentions of adopting the global schema is then lost. Therefore none of the choices for resolving the one-to-none element situation is perfect. Still we would recommend using the Classification category. This would avoid loosing schema grain and lost metadata. Such a decision would open up for sub-local schema cooperation with other LMSs. This would allow for schema extensions with compatibility between the sub-local LMSs. If the global schema should evolve to include these elements, the local schema could convert to the revised schema at that time.

3.6 Many-to-one elements

Many-to-one elements indicate a less grained target schema, allowing less detailed metadata descriptions. We have not found such elements from LO created within this LLMS. There are, however, multiple elements which are not covered within the IEEE LOM schema which could be mapped to the general description element for a many-to-one scenario.

In such a move the different elements would be merged into one element loosing their initial distinct properties; See Table 6. The metadata can then be stored within the schema, though they would not be accessible as individual elements afterwards. An alternative could be performed with local interpretation of the global schema. This would be in conflict with the global metadata schema. Our recommendation is to use the Classification category for these elements.

3.7 Taking advantage of other metadata sources

3.7.1 Automatically creating relations

There are tasks which a LMS can perform without user interaction. This includes updating metadata records with relations not specified by the user. Such relations can be based on:

- Relations between all LOs within the specific course.
- Folders are frequently used to manage LOs into smaller collections, e.g. for creating a compendium. LOs within the same folder can be given their own, additional relations.
- Two-way relations can be created if the LMS have the targeting LO included.
- Some LO types have included links to external sources, e.g. hyperlinks. Discovered links can be used for creating relations.

3.7.2 Creating keywords

The LMS can be an information provider to other algorithms for creating metadata: A course profile, as described in chapter 3.4, can be used indirectly by submitting background information for e.g. a domain ontology algorithm for generating object keywords. This makes the context analysis a basis for content metadata generation.

4 Special challenges regarding files

Our initial studies have shown that 66% of LOs within the LLMS are files. These can currently be described with a single description element. Though files can have much more they can tell.

4.1 Harvestable file element content

When files are created outside of a LMS and without a predefined document template, the LMS has no power to guide and form the content of the files. This being visual properties of the files or their metadata. If the LMS has information of the file format and its metadata schema, it can harvest metadata from such formatted files. Such collectable metadata is shown in Figure 1. Algorithms for file metadata harvesting has been introduced for specific metadata elements in projects including the AMeGA project (Greenberg et al., 2005), the Greenstone Digital Library (Witten et al., 2003) and in LOMGen (Singh et al., 2004).

```
<title>Slide 1</title>
<!--[if gte mso 9]><xml>
 <o:DocumentProperties>
  <o:Author>Lars</o:Author>
  <o:LastAuthor>Lars</o:LastAuthor>
  <o:Revision>3</o:Revision>
  <o:TotalTime>106</o:TotalTime>
  <o:Created>2006-03-08T11:28:10Z</o:C
  <o:LastSaved>2006-03-08T13:14:33Z</o
  <o:Words>208</o:Words>
  <o:PresentationFormat>On-screen Show
  <o:Company>NTNU</o:Company>
  <o:Bytes>20445</o:Bytes>
  <o:Paragraphs>48</o:Paragraphs>
  <o:Slides>12</o:Slides>
  <o:Version>11.6568</o:Version>
 </o:DocumentProperties>
 <o:OfficeDocumentSettings>
  <o:PixelsPerInch>80</o:PixelsPerInch
 </o:OfficeDocumentSettings>
</xml><![endif]-->
```

Figure 1: Metadata collected from a PowerPoint document

Contrary to the other LO types, the file content is not predefined based on the LLMS' LO types. A file can contain a questionnaire, a list of student names or have any other content. When uploading a file to the LLMS, there are no elements available to determine the LO type of the file contents.

File harvestable metadata opens for extensive metadata collection. Since these files where created outside of the LLMS, there are questions regarding the content of extracted metadata elements. One issue is less informative entities: e.g. in Figure 1 the author element has the entity "Lars". This is a less informative element than the full name collectable from the LLMS. Collectable metadata can also include errors which conflicts the file's metadata schema. Our studies have uncovered examples where file metadata elements have been replacement with advertisements.

Other elements can give more descriptive and precise metadata descriptions than elements created within the LLMS. This includes the element for document language; the LLMS do not have a dedicated element for LO language, whereas many text based documents contain registration of the actual language used.

LMSs must be maintained in order to recognize and take advantage of the currently used file formats.

4.2 One-to-none elements

Similar to the LLMS' other LO types; files can contain metadata which are not covered by the global metadata schema. These issues and solutions are equal to the LLMS' LO

types, though the amount of elements with missing global elements can increase. We have discovered missing IEEE LOM elements for a file's number of pages, slides or spreadsheets, paragraphs, lines, words, characters, notes and creator- and producer application. For multimedia content there are missing elements for:

- Image: Resolution (dpi), number of pixels, colour depth
- Sound: Number of channels, bit-rate, actual content playing time
- Multimedia: Frames per second, image and sound metadata

In order to cover these elements lossless within the IEEE LOM schema extensive use of the Classification category would be required.

4.3 Many-to-one elements

When including files as a metadata source, this increases the number of candidate elements sources within the LLMS. Selecting the best candidate element can then be more challenging. For example we want to give a LO the correct title. The title element is specified in the LLMS and in the metadata for many file formats. Many documents can have a harvestable visual title. See the example in Table 8. Here we can choose from four element sources, but IEEE LOM gives room for only one title element. In order to determine the best candidate metadata source, when multiple sources are available, we need techniques to assist in this process.

Table 8: Multiple title sources

LLMS metadata	LLMS title = Exercise nr 2 File metadata title = IT3805 exerc. 2 File name = IT3805exec2 version 1 Visual title = Exercise 2 – Metadata
IEEE LOM metadata	1.2 Title = ?

5 SUMMARY AND FUTURE WORK

Converting a local LMS' metadata schema to a global schema requires extensive information about both the local and global schemas, the elements they contain and the intentions behind each element:

- The local LO types, their properties and how they can be used
- The local setting in which the LOs are created or published
- The "hidden knowledge" not explicitly present within the local schema or the LO, though available through local knowledge of the LMS, the LOs and the local educational system
- Available data sources and their potential metadata element sources, and
- The targeting metadata schema, its available elements and their intended usage.

Within the LLMS there is a potential to create rich IEEE LOM metadata records, where the data collection can be based on multiple data sources. This opens up for creation of descriptive metadata records with many finely grained elements enabling precise LO queries.

The technologies developed as crosswalks for a 2-way metadata transferral between schemas, have shown validity for this project. We have uncovered extensive schema mapping possibilities where:

- Single local elements described multiple IEEE LOM elements
- Local elements without a direct equivalent within the IEEE LOM schema
- Multiple local elements describing a single entity IEEE LOM element
- Reduced reliability caused by LO elements containing error-full metadata.

We have discovered that the file LO type is the prime candidate in order to locate Many-to-one elements. Files have shown to be a less reliable metadata source.

There are unresolved issues regarding how to deal with elements that are not covered by the current IEEE LOM version. Excluding these elements results in lost data. Using the Classification category results in elements not understood by other LMSs and services using the global schema.

In future work we will analyze the content of discovered metadata sources. This includes LO files collected from the LLMS in the Adobe PDF, MS Word, MS PowerPoint, MS Excel and JPEG file formats. We will analyze elements which have shown to contain entities and comparing elements where there are multiple candidate sources. This includes elements for title and author name. We will compare the results between the different file formats and the other LLMS' LO types.

By doing these efforts we will show which metadata sources that are available based on the LLMS, which metadata sources that should be used and which, if any, metadata sources to give priority.

REFERENCES

CETIS, 2004, *UK LOM Core v 0.2*, http://www.cetis.ac.uk/profiles/uklomcore/uklomcore_v0p2_may04.doc

Chan, L. M., Zeng, L. M., 2006, *Metadata Interoperability and Standardization – A Study of Methodology Part I - Achieving Interoperability at the Schema Level*, D-Lib Magazine, June 2006, Volume 12 Number 6, ISSN 1082-9873

eStandard, 2005, *Norsk LOM-profil – NORLOM. Versjon 1.0*, http://www.estandard.no/norlom/v1.0/NORLOM_v1_0_mars_2005.pdf

Greenberg J., Spurgin, K., Crystal, A., Cronquist, M., Wilson, A., 2005, *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project*, UNC School of information and library science, http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

IEEE LTSC, 2005, *IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata*, WG12: Related Materials, http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

Library of Congress, 2001, *Dublin Core/MARC/GILS Crosswalk, Network Development and MARC Standards Office*, 2001-03-12

Singh, A., Boley, H., Bhavsar, V.C., 2004, *LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology*, National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004, http://www.cs.unb.ca/agentmatcher/LOMGen/paper/LOMGen-29Mar2004.ppt

Witten, I. H., Bainbridge, D., Boddie, S., Don, K. J., McPherson, J. R., 2003, *Greenstone Digital Library – Inside Greenstone Collections*, http://www.greenstone.org/docs/inside_greenstone.pdf

Zeng, M. L., Xian L., 2001, *Mapping metadata elements of different format.* E-Libraries 2001, Proceedings, May 15-17, 2001, New York: 91-99. Medford, NJ: Information Today, Inc.

P2: Automatically generating high quality metadata by analyzing the document code of common file types

Published in the Proceedings of JCDL 2009
June 15-19, 2009, ACM

Lars Fredrik Høimyr Edvardsen
Intelligent Communication AS /
Norwegian University of Science and Technology
Kristian Augusts gate 14
NO-0164 Oslo
+47 908 41 765
lars.edvardsen@intelcom.no

Trond Aalberg
Norwegian University of Science and Technology
Sem Sælands vei 7-9
NO-7491 Trondheim
+47 73 59 79 52
trond.aalberg@idi.ntnu.no

Ingeborg Torvik Sølvberg
Norwegian University of Science and Technology
Sem Sælands vei 7-9
NO-7491 Trondheim
+47 73 59 60 27
ingeborg.solvberg@idi.ntnu.no

Hallvard Trætteberg
Norwegian University of Science and Technology
Sem Sælands vei 7-9
NO-7491 Trondheim
+47 73 59 34 43
hal@idi.ntnu.no

ABSTRACT: A major challenge for content management in intranets and other large scale document storage and retrieval services is the generation of high quality metadata. Manual generation of metadata is resource demanding and is often viewed by collection managers and document authors as inefficient use of their time, and there is a desire for other ways to create the needed metadata. Automatic Metadata Generation (AMG) is methods for generating metadata without manual interaction using computer program(s) to interpret the document and possibly the document context. Current AMG research has been limited to collection of similarly formatted documents. The research presented in this paper expands the field of AMG by presenting an approach that is independent of a common visualization scheme; AMG based on document code analysis. This is done by showing AMG possibilities from Latex, Word and PowerPoint documents and how this approach can significantly increase the quality of the generated metadata. This by avoiding common quality reducing factors as missing completeness, low accuracy, logical consistency and coherence and timeliness by giving AMG algorithms direct access to the user specified intellectual content and the file formatting. This research shows how this AMG approach can be combined with other AMG approaches, drawing on their strengths in order to achieve the desired high quality metadata entities.

Categories and Subject Descriptors
H 3.1 [Information Systems] Content Analysis and Indexing – abstracting methods, indexing methods
H 3.7 [Information Systems] Digital Libraries - collection

General Terms: Algorithms, Reliability, Experimentation, Verification.

Keywords: Automatic Metadata Generation, Harvesting, Extraction, Document Code, Metadata Quality, Latex, Word, PowerPoint, PDF, OpenXML.

1. INTRODUCTION

Metadata is commonly used to describe the characteristics of resources. The main purpose of metadata is to support querying and retrieval of relevant content. AMG is based on the observation that information that equals the desired metadata often already is contained in the documents, such as:

- Visual and technical descriptions: E.g. formatting information and the number of visual pages.
- Intellectual content descriptions: User specified textual content. E.g. the document title and author.

The document author has hence directly or indirectly specified the desired content of many metadata elements. Based on this, why should we manually reproduce something which is already available? AMG strive to avoid excessive manual efforts when similar metadata can be generated automatically based on existing data sources [1, 2, 3, 4]. Related AMG research has been focused on three directions for generating metadata:

a) Harvesting the existing metadata that can be found in document files.
b) Generation of keywords based on natural language document analysis.
c) Extraction of content from pre-specified locations of the visually presented document.

As shown in previous work [5], these approaches all have major weaknesses: (a) Harvestable metadata is often faulty due to wrong content used for generating metadata. (b) The method of natural language, full-text analysis only work for keyword and statistical metadata generation. (c) Extraction based on visual characteristics is limited to pre-known, similarly formatted documents. This is since identification of the "right" content for the metadata element can vary visually from document to document. These are limitations which need to be addressed before AMG on larger, less structured document collections can be performed. This research aims to spread knowledge of a fourth approach for generating metadata: By analyzing the file format specific formatting, "the document code", in order to recognize key characteristics regardless of visual characteristics of the document. This approach can be used to identify the author specified document content for generation of high quality metadata without the need for visual identification of document content.

In order to promote local and global sharing and reuse of existing resources published at Norwegian University of Science and Technology (NTNU), this research has desired to label resources with metadata in accordance with the international IEEE LOM metadata

schema standard [6, 7]. This educational metadata standard is extensive, enabling rich and detailed resource descriptions, while being backwards compatible with the Dublin Core schema [8]. The IEEE LOM is also a prime example of a metadata schema which can be difficult to fill out by end users; A process which can take more than an hour to fill out per document by trained uses.

Through analysis of content collected from the Intranet at NTNU, a so-called Learning Management System (LMS) named It's:learning [9], this research has developed ways in which to automatically generate high quality metadata without being a burden on the publishers and document authors. In this paper, the focus is on semantic elements, hence elements which promotes subjective author- or publisher specified content. In order to enable access to the documents' semantic content, there is a need to understand the formatting of the document code correctly and automatically. Hence, there is a need to determine the file format of the document. The following elements have been used to illustrate AMG potentials:

Table 1. Metadata elements examined

Used element name	IEEE LOM	Dublin Core
Format	Technical.Format	Format
Title	General.Title	Title
Language	General.Language	Language
Keywords	General.Keywords	Subject
Description	General.Description	Description
Creator	Lifecycle.Contribute.Role=Creator Lifecycle.Contribute.Entity	Creator

Of the elements above in Table 1, the Title, Language, Keywords, Description and Creator (author) are semantic elements. To specify a Creator, the IEEE LOM uses two elements: One for the role and one element for the creator name. Format is a syntactic element used to determine how to understand the file content.

Chapter 2 presents the state-of-the-art of the field of AMG and defining the term "quality". Chapter 3 describes weaknesses in current AMG efforts when used on such a visually diverse and multi linguistic document collection environment as the NTNU Intranet. Chapter 4 presents how analysis of the document code of common document formats can be used to generate high quality metadata, with a special focus on the Title element. Chapter 5 presents usage of the document code analysis as data source and contributor to other AMG approaches for generating keywords and descriptions. Chapter 6 gives an advice exemplified by the Creator element before Chapter 7 concludes and presents future work.

2. BACKGROUND

2.1. Automatic Metadata Generation

AMG algorithms are sets of rules for the processing of data source(s), identification of desired content, and the collection and storage of data in accordance with a metadata schema. AMG algorithms can use the document itself and the context surrounding the document as data sources. Collecting embedded metadata is known as metadata *harvesting* [2, 10]. The process by which AMG algorithms create metadata that previously has not existed is known as metadata *extraction* [11, 12]. AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that is incorrect for the description of a document. Document content analysis is currently the main approach for generating document specific metadata. Four different approaches are used, as presented in Figure 1:

- **Harvesting of embedded metadata.** This approach uses the embedded metadata created by applications or by the user and stored as part of the document [12, 13, 14, 15, 16, 17]. This approach is vulnerable to generating false metadata if the embedded metadata is incorrect.
- **Extraction based on visual appearance.** This approach uses a content presentation application to create a visual representation of the document before executing rules to extract content based on the visual appearance of the document [18, 19, 20, 21, 22]. This approach is vulnerable to generating false metadata if the documents do not share the visible appearance(s) with which the algorithm has been developed to perform. Hence, such algorithms only perform as desired on pre-known document types.
- **Extraction of metadata based on natural language.** This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed [23, 24, 25, 26, 27, 28]. Such algorithms commonly include collection of unique words and comparisons of the document vocabulary against reference ontology for generating keywords, descriptions and subject classification. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages, document sections in different languages or contains header or footer fields since the text from these fields are presented on every page hence occur frequently.
- **Extraction based on document code analysis.** This approach uses analysis of the code of e.g. a document directly without the need for additional content presentation applications to interpret the document content. This enables full and direct access to the entire document's content. This includes template identification, template content identification and formatting characteristics regardless of visual characteristics, and the language of the intellectual content. Current, popular document formats are binary (e.g. PDF, Word and PowerPoint) or non-standardized (e.g. Word and PowerPoint). This has limited the research based on document codes to HTML documents [28]. With the emergence of new document file formats; this paper will explore the use of the document code on Word and PowerPoint documents.

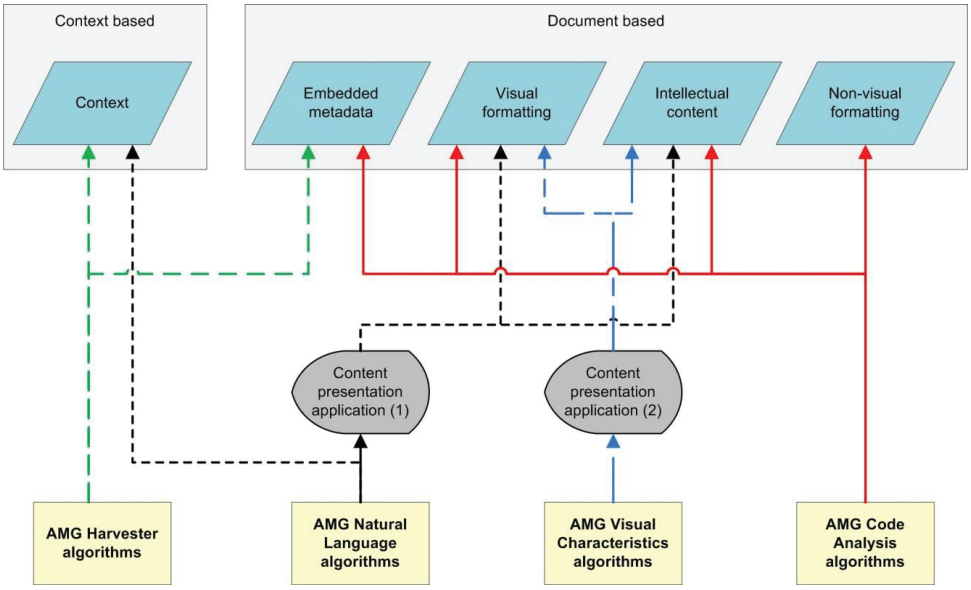


Figure 1. AMG content analysis algorithms and the data sources which they use

2.2. Defining Quality

The research results were evaluated using a framework for measuring "quality" presented in [29]. In the context of this research, this framework categorizes "quality" based on:

1) **Syntax:** Analysis of the document formatting to see that it complies with the document's format standard.
2) **Semantics:** Analysis of the entities presented to see if they are valid and in accordance with the document format's relevant metadata schema.
3) **Pragmatics:** Analysis to determine if the user-interpreted properties are reflected in the metadata.

Additionally, supplemental quality terms were used based on [30]. This framework supplement [29] by including dedicated metadata quality terms for:

1) **Completeness:** Completeness reflects two issues: (1) The use of as many elements as possible; and (2) that the user's desired elements are present in the metadata records.
2) **Accuracy:** The entities should describe the document correctly and factually.
3) **Provenance:** There should be a record of who created the metadata.
4) **Conformance to expectations:** Assumes that the user's expected elements are available.

5) **Logical consistency and coherence:** Logical consistency relates to compliance with the local metadata schema. Coherence relates to whether the elements are made available.
6) **Timeliness:** Timeliness relates to two issues: (1) Currency: when the document changes while the metadata remain unchanged. (2) Lag: when the document is disseminated (distributed) before some or all metadata is knowable or available.
7) **Accessibility:** That the metadata are available to users and understandable to users.

The quality scale is measured subjectively as:

- **Very high:** The dataset can confirm a high degree of correctness.
- **High:** The dataset can confirm a high degree of correctness, although more than a few exceptions were discovered.
- **Undeterminable:** The dataset could not verify either correct or false entities for the given element, so that a conclusion could not be drawn.
- **Low:** Systematic false entities were verified to be present.
- **Very low:** An extensive number of false entities were verified as present in the dataset.

3. FINDING STRUCTURE IN CHAOS

A common strategy for ensuring the existence of metadata is to force publishers to manually specify metadata. At the NTNU Intranet we try to avoid such force and hence automatically generates publishing metadata based on login and session data. However, the user must manually specify a document title. In addition, a dedicated document description can be manually added. The title specified through the Intranet showed similarities with the documents' visible title for less than a tenth of the published documents. Description or Subject metadata were specified for two thirds of the published documents, though less than one thirds of the documents were individually described. These numbers indicate that the users regard creation of these metadata as unnecessary extra work, and is not worthy of providing extra effort into. There is a need for other methods for generating quality metadata. No documents were retrieved from the Intranet with valid Description or Subject metadata, even though common document formats such as PDF, Word and PowerPoint support inclusion of such entities.

At NTNU the lecturers and students themselves decide upon which documents they wish to share through the Intranet. As a result of this free publishing policy, a wide range of lectured subjects and a multitude of physical lecturing areas result in an Intranet document collection with extremely diverse visual appearances. Here you can find everything from highly structured academic papers to the students' presented answers to exercises varying in subjects including medicine, informatics, education and fine art. Figure 2 shows a few examples of the visual diversity of published documents. This research analyzed over 8000 documents from this Intranet, collected from 166 unique courses for analysis. Random selections of from the 3500 stand-alone document files collected from this dataset were used for in-depth analysis. This analysis confirmed that existing AMG algorithms based on harvesting of existing metadata and extraction

based on pre-defined visual characteristic or natural language analysis would not perform in compliance with the quality measures of the University [5].

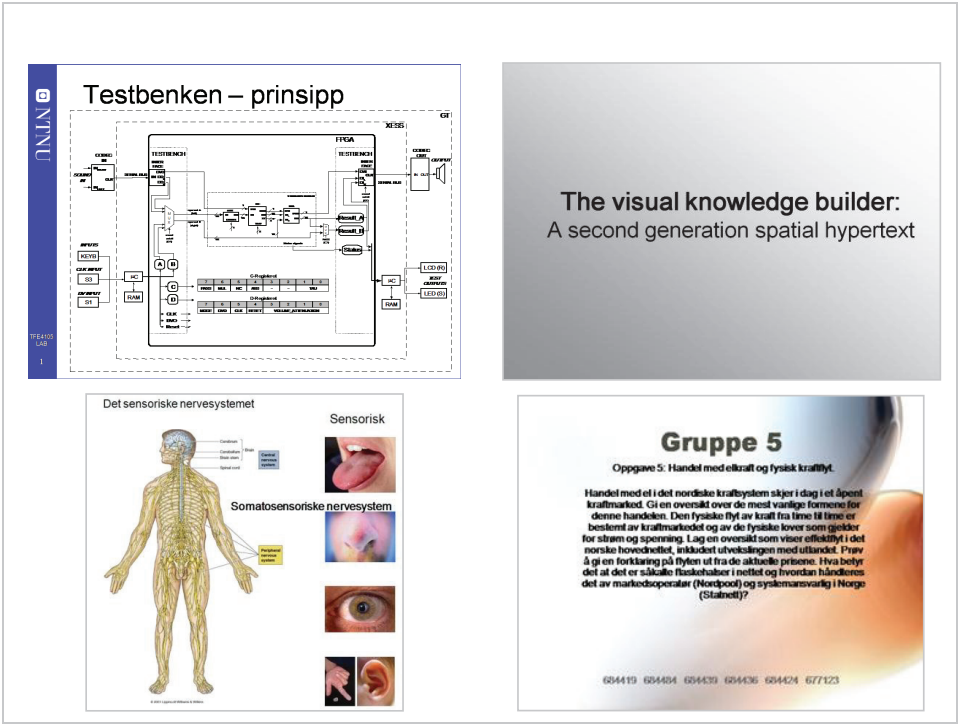


Figure 2. Four examples of published PowerPoint slide shows from the NTNU Intranet.

The low or very low metadata quality was caused by:

- **Harvestable metadata being based on standard entities** not updated entities and false entities due to document conversion. E.g. only 14% of dataset documents contained a higher quality Title element. In regards to PDF documents, this low metadata quality were results of the original document metadata being excluded and replaced by commercial content promoting the converter application. This results in very low metadata quality in terms of semantics & accuracy (describes other content than the intentions of the schema) and pragmatics & logical consistency and coherence (the metadata does not reflect upon the document). MS Office documents were affected negatively by standardized template entities (e.g. Author = NTNU) and timeliness due to metadata not being updated as the document evolves (e.g. the Title-element is populated the first time the document is saved, though since need to be manually updates. This was seldom done in the analyzed dataset). None of the examined

document formats contained meta-metadata and hence all had metadata of low quality in terms provenance. As a result, it was in cases impossible to determine if the metadata were user specified or generated by an application.

- **Extractable metadata based on visual characteristics** not being able to locate the "correct" sections due to extensive and unstructured differences in visual appearances. E.g. one in five Word-documents received a Title element without any resemblance to the visible title. Common issues included low completeness due to titles spanning multiple sentences and accuracy due to wrong content being used as metadata (e.g. the first line of text).
- **Extractable metadata based on natural language**, full-text analysis was not able to distinguish between visual document sections. E.g. the footer and header of documents were mixed up with the documents' intellectual content reducing the pragmatic and conformance to expectations. Hence, words such as "Page" and "NTNU" were often located as keywords even though the intellectual document content did not use these words.

As the harvestable metadata is of questionable quality, AMG need to focus on extraction. Though, at the same time move beyond extraction limited to visual characteristics. [31] showed retrieval of the Title-element from HTML tags and extraction based on the natural language approach by analyzing the HTML code of web-pages. The generated metadata were of less than their desired quality. They concluded that other means of generating the Title element were needed which could combine data sources, in order to generate higher quality metadata.

The file format of other commonly used stand-alone document formats contains extensive descriptions of their content as well. Ninety-one percent of the stand-alone documents uploaded to the NTNU Intranet were in Adobe PDF, MS Word or MS PowerPoint document formats. This research has focused on AMG efforts from these document formats. These similarities based on document format provide extensive amounts of information which can be used for AMG purposes regardless of visual appearance. Using document code analysis as basis for AMG allows for combining AMG approaches, hence being able to use the documents' visual appearance to automatically generate metadata when this is preferable.

4. EXTRACTION OF STANDARDIZED FILE FORMATTING CONTENT

The structure of the document code is determined based on the documents' file format. Each file format has a pre-specified structure with a pre-specified logical consistency and coherence. It is hence of essence to determine the file format correctly in order to gain access to and to understand the document code. The file format can usually be determined based on the file name extension. The syntactical data quality of the examined documents was in general very high with file content in conformance to expectations. Only three documents were retrieved with a document code which did not correspond to the conformance to expectations. These were all corrupted MS PowerPoint files. In addition, several PDF documents were security restricted. Such security restrictions can be a hurdle for AMG efforts, since the AMG algorithms might

not gain access to all or the desired sections of the target document. The result is low syntactic quality due to low content accessibility even though the file formatting is of high syntactic quality.

Documents of a specific file format contain extensive amounts of similarly structured formatting data due to their commonly built file format syntax. This is even though the documents' visual appearance can vary extensively. By building on these similarities, structure can be established to build AMG logic. E.g. Latex documents based on the ACM SIG Proceedings Template [32] share a specific formatting specified in the template. By building AMG algorithms that are adapted to the used document template, extraction of specific sections can be used to generate metadata. Basing the extraction efforts on the document code rather than the visual appearance of the document avoids visual abnormalities to affect the quality of the generated metadata. E.g. from Figure 3 the Title-element can be identified by locating the "\title"-section, while the right number of authors can be uniquely established by retrieving and analyzing the content of the "\numberofauthors"-section. The content of these sections are always up-to-date as long at the sections are used in accordance with the given template. There are hence no timeline issues which could drag the quality down as long as the syntactical quality is high and has high accuracy. Hence, no additional efforts are needed in order to determine and extract the correct number of authors. This avoids the uncertainties which visual recognition based AMG algorithms face.

```
\begin {document}
\title {AUTOMATIC METADATA GENERATION}
\numberofauthors {1}
\author {
\alignauthor Lars Edvardsen\\
\email {lars.edvardsen@intelcom.no}
```

Figure 3. Example Latex document code based on the ACM SIG Proceedings Template.

A substantial challenge for Latex documents is that they are seldom published in their native format. E.g. at the NTNU Intranet contains only a handful of Latex documents. Rather, converted documents are published instead. The converting process dramatically affects the syntax of the original user specified- and file formatting content. This since most conversion processes can remove the non-visual formatting content of the original document and replaces these with the new file formatting. Original content which is not of explicit importance for the documents' visual appearance is often excluded. E.g. the PDF format does not support inclusion of the non-visual formatting data from Latex or other document formats. The document quality in terms of completeness of the original file is substantially lowered. Instead, new formatting data is included to provide the visible characteristics of the document. This exclusion of original document content limits the extraction possibilities which are based directly on the document code.

Other document types are more commonly published in their original file format. This includes MS Office documents. As a result, all the original documents' content is

available for AMG efforts. The quality of these files is hence higher in terms of syntax and completeness than converted documents. However, the quality in terms of accessibility of MS Office documents has been very low since these formats have been largely unavailable for research due to its proprietary file formats. The introduction of OpenXML-based documents in MS Office 2007 application suite has enabled insight into the file format content. By losslessly converting existing documents collected from the NTNU Intranet into OpenXML, this research has increased the quality of accessibility, which has enabled examination of the actual content of Office files for AMG purposes.

Examination of this document code revealed this as an exceptional data source for generating high quality document metadata. This since it enables access to all the user-specified document content plus all the file format type specific formatting. This enable generation of AMG algorithms which only generates entities of the desired content has been explicitly specified in the document code. If the desired content is explicitly specified then other AMG efforts can be executed. Some of the characteristics of MS Office documents are results of interpretations from the document presentation application. E.g. the exact visual appearance of Word documents is not stored. Rather, the visual appearance is rendered each time the document is opened by the document presentation application.

The OpenXML code in Figure 4 presents Word document code based on the ACM SIG Proceedings Template [32]. This template contains sections for "**Paper-Title**", "**Author**" and "**E-Mail**". These are template specific sections which are presented as "styles" in the application graphical user interface, see Figure 5. Other templates may contain other styles. Office documents do not contain data for logical consistency, such as \numberofauthors from the Latex template. This avoids logical consistency quality issues, though increases the uncertainties regarding the actual content of the document.

```
<w:pPr>
  <w:pStyle w:val="Paper-Title" />
</w:pPr>
<w:r w:rsidRPr="00E82FB1">
  <w:t>AUTOMATIC METADATA GENERATION</w:t>
</w:r>
<w:pPr>
  <w:pStyle w:val="Author" />
</w:rPr>
<w:t>Lars Edvardsen </w:t>
<w:pPr>
  <w:pStyle w:val="E-Mail" />
<w:rPr>
<w:t>Lars.Edvardsen@intelcom.no</w:t>
```

Figure 4. Example OpenXML document code based on the ACM SIG Proceedings Template.

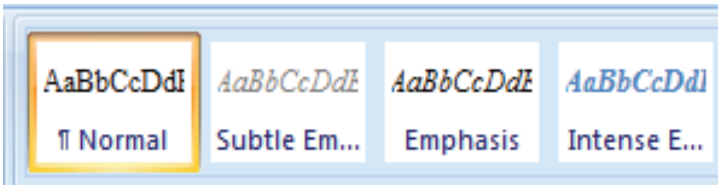


Figure 5. Style "normal" in MS Word 2007.

Each of these styles has a specific visual formatting which is specified as part of the template syntax. If the document author uses these specified style section in accordance with the given template semantics, then the document's visual appearance will be in line with the template standard. This is since the document presentation application uses the styles and the related, specified intellectual content to generate the visual appearance of the document. Hence the quality in terms of semantics and pragmatics will be high. Only the intellectual content which is formatted based on the used template section is registered as being based on this section. A consequence of this is that the styles and the style formatted intellectual content can be individually identified. Using this information for AMG purposes, it is possible to consistently extract document content based on a specific style and use these for generation of metadata suitable for the specified style, increasing the quality in terms of accuracy. If the user does not use the specified style, then no sections of the document's intellectual content contain the specified style. E.g. if the user does not explicitly specify content to the specified "**Paper-Title**" section, then there is no such section present as part of the intellectual content section of the document code. Based on this, AMG algorithms based on document code analysis should only generate metadata if the desired formatted section is not present (user specified). If the user does not follow the document guidelines presented though the template, the quality in terms of semantics and pragmatics is affected negatively.

Each template can have its own styles and can use each stile in a unique way. It is hence of importance to identify which template which was the basis for the document. Identification of the used template can be performed by retrieving the template name which is harvestable of high quality from the document metadata present in most documents. Analysis of thousands of documents from the NTNU Intranet revealed that all the Word documents had a registered template. The standard template "Normal.dot" were the basis for over 95% of these documents. This blank document template contain sections for e.g. title and headings, though not author, abstract, description or keywords as the ACM template does. The section names also differ, e.g. the "Normal.dot" specifies the title as either "**Title**" or "**ctrTitle**", while the ACM template specifies its title as "**Paper-Title**". Only document content which the author explicitly specifies to be of a specific type is formatted in accordance with this style. In-depth analysis revealed that only 3% of Word documents contained Title-styled intellectual content of Word documents. This stands in vast contrast to the 82% of PowerPoint slide

shows which contained such data. By using the document code as the primary data source for harvesting and extraction algorithms, this research increased the correctness rate of the generated metadata to 91% and 97% for Word and PowerPoint document respectfully with only 3% of the entities generated being false [5]. The approach of combining AMG efforts is explored in the next chapter.

The high number of usage of the Title-styled section of PowerPoint slide shows is a result of most templates, including the default blank template, visually promotes using these sections. Hence many users use these sections to get a slide show which complies with the theme of the used template while increasing the document quality in terms of completeness [4]. The authors are hence encouraged to contribute with specific content at the style formatted sections, which can be uniquely identified and used for AMG efforts regardless of the visual appearance of the slide show. Figure 6 present one of many observed used templates which promote a title and a subtitle section on the front page, alias the "Title slide". By filling these sections out, the intended visual appearance of the document is accomplished.

Figure 6. Template with title and sub-title sections.

5. RETRIEVING OTHER FORMATTING CONTENT

5.1. Generating descriptions

In accordance with the IEEE LOM [7] and Dublin Core [8], the metadata element "Description" can be populated with entities which include a table of content. Table of

contents can be automatically generated by analyzing the document code for retrieval of all the sections of the documents' intellectual content formatted as a header, such as w:val="**Heading1**". It is also possible to retrieve the Table of content directly if such is created within the document. Though, this potential data source can have timeliness issues deterring the metadata quality. This is due to the Table of content is not continuously updated, rather updated when manually specifying or when printing the document. Other content which is described as a Caption can also be retrieved for generating Table of tables, figures, illustrations etc. Figure 7 present the actual OpenXML code which is used for creating the Caption of Figure 7.

```
<w:pStyle w:val="Caption" />
<w:t xml:space="preserve">Figure</w:t>
<w:fldSimple w:instr="SEQ Figure \* ARABIC">
  <w:t>7</w:t>
</w:fldSimple>
<w:t>. A selection of OpenXML document code to generate
     this specific caption. </w:t>
```

Figure 7. A selection of OpenXML document code to generate this specific Caption.

5.2 Using code analysis as basis for other AMG efforts

A range of metadata cannot be explicitly specified using the document alone. In many cases there is a need to include logic from other AMG approaches in order to generate metadata of the desired quality and quantity.

5.2.1 Extraction of Abstract

One such entity is the Description element as a summary or abstract. No documents retrieved from the NTNU Intranet contained sections which were style formatted as "Summary" or "Abstract". Though, a larger selection of documents had an abstract paragraph located beneath the author information on the first page. This abstract session was usually marked with the header text "Abstract" and consisted until the next header. Descriptions such as the one above are the backbone for AMG algorithms based on visual appearance. Key challenges for such algorithms are to extract all the desired intellectual content (completeness) without extracting unwanted content (accuracy).

Analysis of the document code can contribute with information to improve the algorithms' capability to generate high quality metadata in terms of completeness and accuracy by contributing to the visual appearance algorithm with facts regarding the content of the document. Analysis of the document code enables unique identification of the Abstract header (commonly formatted as a style header, formatted with Capital letters, bold letters or using a larger font than the continuing text). The source code is formatted as a single column regardless of the number of column present in the visible document. Retrieving the Abstract content can hence be performed by extracting all text

between the Abstract header and the next header located in the source code. This avoids completeness and accuracy issue challenges, such as extraction of the right order of paragraphs and extraction of content from one or more columns dependent upon the visual presentation of the document. As a result, the pragmatic quality of the metadata can increase in terms of completeness and accuracy.

5.2.2. Extraction of Keywords and Subject classification

Other document content which were not experienced explicitly present in the source code of the average document were keywords and subject classification. It would be possible to extract such content using the same logical approach as extracting the Abstract section from documents with such information present, such as documents based on the ACM SIG Proceedings Template [32]. Unfortunately, most publications on the NTNU Intranet are not based on this or similar templates. Other efforts are hence needed to generate the desired metadata.

Within the field of AMG efforts based on natural language, many algorithms have been developed to generate metadata such as Keywords and Subject classification [23, 24, 25, 26, 27, 28]. Determining the language of the documents' intellectual content is of absolute importance for these algorithms in order to generate quality metadata when operating in a multi linguistic environment. Current AMG efforts based on natural language have avoided this issue by using a dataset of documents with intellectual content in a single language, usually English. This assumption cannot be used on documents published on the NTNU Intranet as this is a multi linguistic user environment.

Current AMG algorithms based on natural language are commonly based on the frequency of unique words used in the intellectual content and comparisons of this vocabulary against reference ontology. Based on this, different AMG algorithms have different approaches to selecting the most frequent words and counts of the most uncommon words for generating keywords and subject classification. General purpose "stop words" are commonly removed. Such stop words commonly include "I", "am" and "and" from documents in English. In Norwegian these same stop words would be "jeg", "er" and "og". If a document in Norwegian were to be analyzed based on an English ontology and stop words, then the generated metadata would be of very low quality as the whole content of the document is misunderstood by the AMG algorithm. It is hence of absolute importance to determine the language of the intellectual content for each document and even individual words, sentences and sections. This in order to allow the AMG algorithm to remove the correct set of stop words and use the right language ontology in order to generate metadata.

Analysis of the document code can reveal data which is of importance for AMG algorithms based on natural language in a multi linguistic document environment. A range of applications automatically analyze the document content in order to determine the language of their intellectual content. Applications, such as the MS Office application suite do this and use the functionality to allow its spelling and grammar

checks to perform optimally. The results of this analysis is stored as part of the documents' intellectual content as language tags. See the example in Figure 8.

By analyzing the document code, it is hence possible to review which language each sentence or section were registered as, if there were grammatical faults, if there were false spellings or even if false spelling were ignored. Using this information, it is possible to distinguish between the intellectual content presented in each section and execute AMG algorithms based on natural language adapted to the specific language at hand.

This research's dataset contained Word documents and PowerPoint slide shows which had intellectual content registered as Norwegian, New Norwegian, Danish, Swedish, German, British English, US English, Australian English, Canadian English, Spanish, Portuguese, French and Greek. Several of these documents were multi linguistic.

```
<w:p w:rsidR="006917FF" w:rsidRPr="006917FF">
  <w:r w:rsidRPr="006917FF">
    <w:rPr><w:lang w:val="en-US" /></w:rPr>
    <w:t>This is in English.</w:t>
  </w:r>
</w:p>
<w:p w:rsidR="00EA7686" w:rsidRPr="00EA7686">
  <w:r w:rsidRPr="00EA7686">
    <w:rPr><w:lang w:val="nb-NO" /></w:rPr>
    <w:t>Mens dette er på norsk, bokmål.</w:t>
  </w:r>
</w:p>
```

Figure 8. A selection of OpenXML document code to which specifies the language and intellectual content of each document section.

The accuracy of these language tags was high, though not flawless. The language determining algorithm of the MS Office application suite did show lower accuracy when single words, short or incomplete sentences were present. In these cases, the words and phrases in question were recognized as misspelled in the applications' graphical user interface, and hence were indicated as misspelled in the document code. This functionality can be used to exclude misspelled content and hence avoid intellectual content based on a non- recognized language(s).

Another troublesome issue for AMG algorithms based on natural language is headers, footers and master slide show content. The data from these sections can be present on every page of the document and hence be statistically frequently present in the document. This does not mean that the content of these sections is preferred to be included in the document analysis. The content of these sections are stored as separate sub-files of each OpenXML document. The intellectual content of the Word documents are store in another sub-file, while each PowerPoint slide is stored as an individual sub-

file. It is hence possible to easily distinguish between the user specified intellectual content and the content from the other content sections. Based on this, AMG algorithms based on natural language can select which data source that is preferred used. This can further increase the completeness and accuracy of the automatically generated metadata.

A practical example of an AMG algorithm hierarchy which combines AMG approaches in order to generate high quality Title entities from the same dataset is available as previous work [5]. Here the style formatted title is first attempted extracted. If no such content is retrieved then extraction based on visual appearance is performed. If this too does not result in a valid title, then the harvestable metadata is examined before the file name can be used as a last option. This hierarchy of AMG algorithms generated very high quality Title entities, with only 3% false entities.

6 A WORD OF CAUTION – AVOID GETTING BLINDED BY THE DOCUMENT CODE

Not all document content is strictly formatted in accordance with the given template. This can be a result of the desired style sections not being present or the users' desire to present the document with an alternative visual appearance. This research has experienced that users regard the visual presence of the resulting document as fare more important than the non-visual formatting. If the desired sections are not present in the presented template, other sections can be used for this purpose. An example of this is the sub-title style formatted section which was present in half of PowerPoint slide shows.

In seven out of ten registrations, the creator (author) information was found within this section. These sections were visually formatted in a variety of ways and contained a range of different data, such as subtitles, dates, course descriptions and creator information in a multitude of different orders. Though, less than a tenth of these "Sub-title" sections contained exclusively creator information. The variety in regards to content types and visual formatting makes extraction efforts from this section reliant upon identification of user and organization names, among other text. Deciding upon how to use the available document sections can hence present itself as a challenge. Local knowledge of how the specific (sub-) collection of documents is actually used will enable generation of metadata entities of higher quality due to local adaptation. Such adaptability can be based on e.g. a specific local user, user type, department or organization.

7 CONCLUSION AND FUTURE WORK

AMG algorithms base their efforts on systematic and consistent properties of the documents at hand in order to generate quality metadata in accordance with pre-defined metadata schema(s). AMG algorithms need to find common structures in which to base their efforts, even if the dataset is not homogenous. Recognition of the most correct and

most desirable document properties is the basis for automatic generation of high quality metadata.

The currently used AMG approaches have all strength and weaknesses. Retrieval of existing, harvestable metadata can be the simplest to perform, though these metadata are commonly faced with semantic and pragmatic challenges as a result of low completeness, little accuracy, provenance, logical consistency and coherence and timeliness. AMG algorithms based on visual characteristics can generate extensive amounts of metadata; though can be easily fooled by the visual appearance of the document. In a less structured document publishing environment this lack of visual similarity can lower the semantic and pragmatic quality of the generated metadata substantially. AMG algorithms based on natural language can generate high quality keyword and subject metadata, though are vulnerable for documents in multiple languages or documents of another language then their available ontology.

Even though it can look like there is no structure in a collection of documents, there is often underlying structure based on the file format and common document templates. This paper has showed that analysis of the document code enables insight into the document content and how the common structure of the document codes can used for AMG purposes regardless of the documents' visible appearance. This document code enable direct access to the user specified intellectual content and the style formatting which describes the intellectual content while avoiding undesired content fields, such as headers or footers. This can increase the semantic and pragmatic quality of metadata significantly while avoiding issues caused by low completeness, accuracy, provenance and timeliness. AMG algorithms based on the document code can be combined with other AMG algorithms and provide these algorithms with data sources which enable them to generate higher quality metadata in terms of completeness and accuracy. This paper has demonstrated this potentials by generating high quality semantic entities from a highly diverse and multi linguistic document collection. This resulted in high quality Title and Language entities and made a valuable starting point for generation of Description, Keywords and Subject entities. The generated Creator entities were however of questionable or low semantic quality.

The major bottleneck for examination of the document code has been the syntax. Each file format has a pre-specified structure with a pre-specified logical consistency and coherence. This paper experienced the syntactical quality of files to be high. Though, due to proprietary document formats only the few selected have been able to review the exact content of the commonly used document formats of MS Office documents. This is currently changing as more open source document formats are emerging as viable alternative document formats. The MS Office 2007 based OpenXML document formats have been used for this research to illustrate the potentials of this data source. By basing AMG algorithm efforts on known document format characteristics of these document formats, this research has shown possibilities to retrieve extensive amounts of user specified content usable for generating higher quality metadata. In related work this research has demonstrated how AMG approaches can be combined aiming for automatically generating metadata with as high quality metadata as possible for all documents in the dataset.

The AMG research field is still young and much remains unexplored. At the same time the use of digital documents is increasing dramatically, which offers the potential for extensive research efforts in the years to come. Future work should include (1) Exploring the possibilities for practical experiments using AMG technologies on large document collections; (2) Further evaluation of automatically generated entities which are commonly not explicitly expressed using styles, such as the Creator elements; (3) Research on the use of multi-linguistic documents in generating of semantic metadata using natural language approaches; (4) Usage of data from a controlled user environment as an additional data source in order to automatically generate metadata; (5) Analysis of the similarities between Latex templates in order to generate generic AMG algorithms based on the document code.

8 REFERENCES

[1] Cardinaels, K., Meire, M. and Duval, E. 2005. Automating metadata generation: the simple indexing interface. Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, pp.548-556, ISBN:1-59593-046-9

[2] Greenberg, J. 2004. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging, 6(4): 59-82.

[3] Meire, M., Ochoa, X. and Duval, E. 2007. SAmgI: Automatic Metadata Generation v2.0. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, pp. 1195-1204, Chesapeake, VA: AACE

[4] Duval, E. and Hodgins, W. 2004. Making metadata go away: Hiding everything but the benefits. Keynote address at DC-2004, Shanghai, China

[5] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T., Trætteberg, H. 2009. Using the structural content of documents to automatically generate quality metadata. Webist 2009, March 23-26, 2009. Springer

[6] Edvardsen, L.F.H., Sølvberg, I.T. 2007. Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment. Webist 2007, March 3-6, 2007. Springer

[7] IEEE LTSC, 2005. IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata, WG12: Related Materials, http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

[8] DCMI, 2008. Dublin Core Metadata Element Set, Version 1.1. Dublin Core Metadata Initiative, http://dublincore.org/documents/dces/

[9] It's learning. 2009. It's learning. http://www.itslearning.com

[10] Open Archives Initiative. 2004 Protocol for Metadata Harvesting – v.2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html

[11] Seymore, K., McCallum, A. and Rosenfeld, R. 1999. Learning hidden Markov model structure for information extraction. Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction, pages 37-42, 1999.

[12] Greenstone. 2007. Source only distribution. http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected)
[13] Bird, K. and the Jorum Team. 2006. Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems. 31st March 2006, Version 1.0, Final, Signed off by JISC and Intrallect July 2006.
[14] Google. 2009. Google. http://www.google.com
[15] Scirus. 2009. Scirus – for scientific information. http://www.scirus.com
[16] Yahoo. 2009. Yahoo!, http://www.yahoo.com
[17] Singh, A., Boley, H. and Bhavsar, V.C. 2004. LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology. National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004
[18] Giuffrida, G., Shek, E. C. and Yang, J. 2000. Knowledge-Based Metadata Extraction from PostScript Files. Digital Libraries, San Antonio, Tx, 2000 ACM 1-581 13-231-X/00/0006
[19] Kawtrakul A. and Yingsaeree C. 2005. A Unified Framework for Automatic Metadata Extraction from Electronic Document. Proceedings of IADLC2005 (25-26 August 2005), pp. 71-77.
[20] Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. 2007. Automated Template-Based Metadata Extraction Architecture. ICADL 2007.
[21] Li, H., Cao, Y., Xu, J., Hu, Y., Li, S. and Meyerzon, D. 2005. A new approach to intranet search based on information extraction. Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, Pages: 460 – 468, ISBN:1-59593-140-6, ACM New York, NY, USA.
[22] Liu, Y., Bai, K., Mitra, P, and Giles, C.L. 2007. TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries. JCDL'07, June 18–23, 2007, Vancouver, Canada, ACM 978-1-59593-644-8/07/0006
[23] Boguraev, B. and Neff, M. 2000. Lexical Cohesion, Discourse Segmentation and Document Summarization. RIAO.
[24] LOMGen. 2006. LOMGen. http://www.cs.unb.ca/agentmatcher/LOMGen.html
[25] Greenberg J., Spurgin, K., Crystal, A., Cronquist, M. and Wilson, A. 2005. Final Report for the AMeGA (Automatic Metadata Generation Applications) Project. UNC School of information and library science.
[26] Li, Y., Dorai, C. and Farrell, R. 2005. Creating MAGIC: system for generating learning object metadata for instructional content. Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, pp.367-370, ISBN:1-59593-044-2
[27] Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekema, A., McCracken, N.J., Silverstein, J. and Sutton, S.A. 2002. Automatic metadata generation and evaluation. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11–15, Tampere, Finland, ACM Press, New York, pp.401–402.
[28] Jenkins, C. and Inman, D. 2001. Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF. 0-7695-1022-1/01, 2001 IEEE

[29] Lindland, O.I., Sindre, G., Sølvberg, A. 1994. Understanding Quality in Conceptual Modeling. IEEE Software, march 1994, Volume: 11, Issue: 2, pp. 42-49, ISSN: 0740-7459, DOI: 10.1109/52.268955

[30] Bruce, T.R. and Hillmann, D.I. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions, In Metadata in Practice, D. Hillmann & E Westbrooks, eds., ISSN: 0-8389-0882-9

[31] Xue, Y., Hu, Y., Xin, G., Song, R., Shi, S., Cao, Y., Lin, C-Y., Li. H. (2007), "Web page title extraction and its application", Information Processing & Management, Volume 43, Issue 5, September 2007, Pages 1332-1347.

[32] ACM. 2009. ACM SIG Proceedings Templates, http://www.acm.org/sigs/publications/proceedings-templates

P3: Using the structural content of documents to automatically generate quality metadata

Published in the Proceedings of WEBIST 2009
March 23-26, pp. 354-363, 2009, ISBN: 978-989-8111-83-8, ACM

Lars Fredrik Høimyr Edvardsen

Intelligent Communication AS, Kristian Augusts gate 14, Oslo, Norway
Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælands vei 7-9, Trondheim, Norway
Lars.Edvardsen@intelcom.no

Ingeborg Torvik Sølvberg, Trond Aalberg, Hallvard Trætteberg

Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælands vei 7-9, Trondheim, Norway
Ingeborg.Solvberg@idi.ntnu.no, Trond.Aalberg@idi.ntnu.no, Hal@idi.ntnu.no

Keywords: Automatic Metadata Generation, Extraction, Metadata Quality, Word, PowerPoint, PDF, OpenXML.

Abstract: Giving search engines access to high quality document metadata is crucial for efficient document retrieval efforts on the Internet and on corporate Intranets. Presence of such metadata is currently sparsely present. This paper presents how the structural content of document files can be used for Automatic Metadata Generation (AMG) efforts, basing efforts directly on the documents' content (code) and enabling effective usage of combinations of AMG algorithms for additional harvesting and extraction efforts. This enables usage of AMG efforts to generate high quality metadata in terms of syntax, semantics and pragmatics, from non-homogenous data sources in terms of visual characteristics and language of their intellectual content.

1 INTRODUCTION

Metadata are used to describe the key properties of documents and are normally created by individuals based on a pre-defined metadata schema. The process of manually creating metadata is time consuming and can introduce inconsistencies. These issues can be reduced or avoided by enabling applications to generate metadata instead of or, as a supplement to, manual metadata actions. Such technologies are known as Automatic Metadata Generation (AMG) (Cardinaels et al., 2005; Greenberg, 2004;

Meire et al., 2007). AMG algorithms depend upon data consistency and correct data to generate high quality metadata.

Current AMG efforts are closely related to specific collections of documents with similar visual characteristics and intellectual content based on the same natural language: Boguraev & Neff (2000), Giuffrida et al. (2000) and Seymore et al. (1999) extracts metadata based on highly structured conference-, journal or newspaper template formats. Flynn et al. (2007) automates the document type characteristics before performing visual characteristic AMG efforts, though were still dependent upon recognition of specific visual characteristics. Commonly used document creation applications (content creation software), such as Microsoft (MS) Word, MS PowerPoint and Adobe Distiller, use AMG to generate embedded document metadata, but their quality vary extensively. These data are stored in the document code along with other descriptions of visual and non-visual content.

```
<html>
<head>
  <title>Metadata challenges</title>
</head>
<body lang=EN-US><table>
  <tr><td>Exciting paper on metadata challenges</td></tr>
  <tr><td>
    <p class=Author align=center>
      Lars F. H. Edvardsen and Ingeborg T. Sølvberg</p>
  </td></tr>
</table></body>
</html>
```

Figure 1: The "document code" of a HTML document.

AMG efforts need to generate high quality metadata regardless of visual characteristics and from multi-linguistic documents. This is best undertaken by using the best available algorithm(s) for the specific document, and by using its most desired data sources. The goal of this research was to find methods to automatically generate metadata from non-homogeneous document collections. Basing AMG efforts around document code analysis can enable detailed, structured and correct metadata from non-homogeneous documents. To achieve the research goal, the following questions were answered: (1) What is the quality of automatically generated document content (embedded metadata and document formatting)? (2) Can AMG approaches be combined or selectively used on a document-by-document basis?

Chapter 2 presents AMG basics, while Chapter 3 presents the research approach. Chapters 4, 5 and 6 present the research results. Chapter 7 evaluates the research, with conclusions presented in Chapter 8.

2 AUTOMATIC METADATA GENERATION

AMG algorithms are sets of rules that enable access to data source(s), identification of desired content, collection of these data and storage of them in accordance with metadata schema(s). AMG algorithms can use the document itself and the context surrounding the document as data sources. Collecting embedded metadata is known as metadata *harvesting* (Greenberg, 2004; Open Archives Initiative, 2004). The process by which AMG algorithms create metadata that has previously not existed is known as metadata *extraction* (Seymore et al., 1999; Greenstone, 2007). AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that does not reflect the document. Document content analysis is currently the main approach for generating document specific metadata. Four different approaches are used:

- **Harvesting of embedded metadata.** This approach uses the embedded metadata created by applications or by the user and stored as part of the document (Greenstone, 2007; Bird and the Jorum Team, 2006; Google, 2009; Scirus, 2009; Yahoo, 2009). This approach is vulnerable to generating false metadata if the data sources do not contain high quality metadata.
- **Extraction based on visual appearance.** This approach uses a special content presentation application to generate a visual representation of the document before executing rules to extract content based on the visual appearance of the document (Giuffrida et al., 2000; Kawtrakul and Yingsaeree, 2005; Flynn et al., 2007; Li et al., 2005a; Liu et al., 2007). This approach is vulnerable to generating false metadata if the documents do not share the visible appearance(s) with which the algorithm has been developed to perform. Hence, such algorithms only perform as desired on pre-known document types.
- **Extraction of metadata based on natural language.** This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed (Boguraev and Neff., 2000; LOMGen, 2006; Greenberg et al., 2005; Li et al., 2005b; Liddy et al., 2002; Jenkins and Inman, 2001). Such algorithms commonly include collection of unique words and comparisons of the document vocabulary against a reference ontology for generating keywords, descriptions and subject classification. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages or document sections in different languages.
- **Extraction based on document code analysis.** This approach uses the document code directly without the need for additional content presentation applications to interpret the document content. This enables full and direct access to the entire document's content. This includes template identification, template content identification and formatting characteristics regardless of visual characteristics, and the language of the intellectual content. Current, popular document formats are binary (e.g. PDF, Word and PowerPoint) or non-standardized (e.g. Word and PowerPoint). This has limited the research based on document code analysis to HTML documents (Jenkins and Inman, 2001). With the emergence of new document file formats; this research will explore the use of document code analysis on Word and PowerPoint documents.

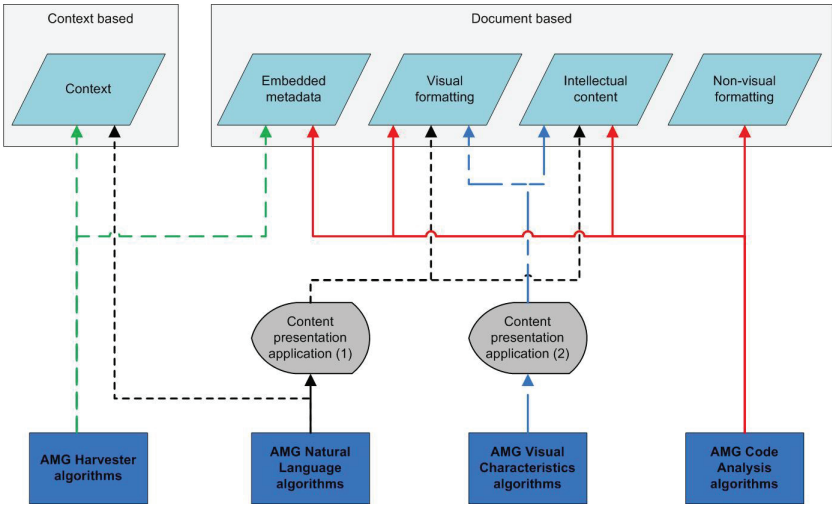


Figure 2: AMG content analysis algorithms and the data sources which they use

3 RESEARCH SETUP

This research needed to base its efforts on documents with diverse visual and intellectual content. These documents were analyzed in regards to their document contents and in regards to generation of metadata. The results of these analysis' were evaluated using an existing framework for measuring "quality".

The Learning Management System (LMS) "It's learning" (It's learning, 2009), which is used by the Norwegian University of Science and Technology, has been used for this project. This LMS allows lecturers and students to publish documents without restrictions regarding document types and visual characteristics, though requiring a user-specified title for each document stored as part of the LMS, not in the files. The LMS automatically generates metadata regarding the publisher based on the logged-in user's user name and gives a timestamp regarding publishing date. This project gained access to 166 distinct courses covering a multitude of subjects, including medicine, linguistics, education and fine art. Here the users published documents without changing any of its characteristics and without restrictions regarding document type or visual characteristics. Over 3500 unique, stand-alone document files were retrieved from these courses.

This project conducted qualitative analyses in order to fine-tune its efforts and gain experience before a more extensive qualitative analysis. For the qualitative analysis, random selections of documents were conducted for in-depth analysis. Ninety-one percent of the stand-alone documents uploaded to the LMS were in PDF, Word or PowerPoint document formats. The qualitative analyses are consequently concentrated on these file formats. The content of the MS Office documents (Word and PowerPoint) was explored by lossless converting them into MS Office 2007 Open XML document

formats using the MS Office 2007 application suite. This conversion process was verified lossless by using third-party software for document content comparisons. The exception is the "Last saved date" metadata elements which were changed. Selected document types are frequently converted before being published, e.g. from Word to PDF document formats. This affects the document content and hence increases the vulnerability to generation of false metadata: (1) Content can be added, altered or removed; non-visible formatting data is commonly discarded. (2) The converted document can contain metadata that reflects the converted document but not the original. (3) Documents can be subject to security restrictions, which prevent AMG algorithms from accessing their content.

The research results were evaluated using a framework for measuring "quality" presented by Lindland et al. (1994). This framework categorizes "quality" based on (1) Syntax, (2) Semantics and (3) Pragmatics. Additionally, supplemental quality terms were used based on Bruce and Hillmann (2004) by including dedicated metadata quality terms for completeness, accuracy and provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility. The IEEE Learning Object Metadata (LOM) (IEEE LTSC, 2005) schema was used to generate a common vocabulary and to define the content of specific elements and their valid value spaces. However, this research is not restricted to this specific schema.

4 QUANTITATIVE ANALYSIS

The LMS shows extensive varieties in regards to published documents, as all documents are accepted for publication. This research found 41 document file formats, a range in content types (texts, spread-sheets, presentations etc.), content qualities (from informal notes to papers) and intellectual content in a multitude of different languages. The documents have a diverse visual appearance, ranging from being based on predefined official administrative templates used by university employees, to documents without structure created by students on private computers. The following embedded metadata elements (and their synonyms) from these documents have been analyzed:

The "Date" elements: All Word and PowerPoint documents and 91% of PDF documents contained embedded date metadata. However, less than a handful of documents contained visible date content against which an evaluation could be performed. All the embedded "Date" elements were based on the timer (clock) of the user's local computer. There was no information stored as part of the document or from the LMS that could verify that this timer was correct when the metadata was generated. Therefore, the correctness of these entities cannot be determined, although a few elements could be confirmed as being false, because the entities indicated that they were modified before being created or that they were published before being created or last saved. This confirms that "Date" elements cannot be fully trusted.

The "Creator" element: All Word and PowerPoint documents and 76% of PDF documents contained a "Creator" (or "Author") element. These elements are commonly automatically generated by applications using software license user names and default

values. Only 46% of PDF, 22% of Word and 30% of PowerPoint documents contained visible author information, making validation of these entities challenging.

The "Template" element: Ninety-five percent of Word documents were based on the blank default template, which is without any visible content. Eighty-two percent of PowerPoint documents were based on the application's default template "normal.pot" which contains visible "Title" and "Sub-title" sections. These sections are identifiable and retrievable though analysis of the document code. This template information is discarded when the original documents are converted to PDF documents.

The "Title" element: All Word and PowerPoint documents and 84% of PDF documents contained a "Title" element. These elements are commonly automatically generated by applications the first time the document is stored. The documents' visible title and the embedded metadata "Title" were identical for only 14% of the documents. This indicates that the visible titles were changed when the documents were resaved or that the AMG algorithms used generated false entities.

The "Description", "Subject" and "Keywords" elements: Just 0.1% of the Word and 1% of PDF documents contained a "Description" element. Most of these entities were valid. No PowerPoint documents contained valid "Description" elements. One percent of PDF documents contained a "Subject" entity, though only one-eighth of these entities were valid. No documents contained a valid "Keywords" entity.

The "Language" element: No documents contained metadata regarding the language of the document's intellectual content.

The quantitative analysis was used as basis for the further efforts of the qualitative analysis. There is no more data in the dataset to determine the correctness of the "Date" elements. Further analysis has therefore not been undertaken. Further analysis is presented in Chapters 5.1 and 5.2 regarding the "Creator" and "Title" elements. These efforts use the "Template" entities. The uncommon, but valid use of the "Description", "Subject" and "Keywords" elements show the need for AMG algorithms based on natural language. In a multi-linguistic environment, these algorithms are dependent on document and document section language information. This is discussed in Chapter 5.3.

5 QUALITATIVE ANALYSIS

5.1 Generating "Creator" elements

This chapter analyses embedded "Creator" entities of common document formats and AMG approaches for generating such entities. For this analysis, 300 PDF, Word and PowerPoint documents were selected at random. Visual data to verify element content were present in only a limited way, which increased uncertainties and our ability to draw conclusions regarding the embedded metadata and the extracted metadata. Word and PowerPoint documents can have embed "Author" and "Last author" elements. PDF documents can embed a general "Author" element and an Extensible Metadata Platform (XMP) section with "DC.Creator" and "XAP.Author" elements. The entities presented

in the XMP section contained a number of character errors, with characters being added, removed or replaced. All these entities were also found in the general element section but without the issues described above. These elements could therefore be used exclusively without losing data. The majority of documents contained author or organization names in their embedded metadata, though only a fraction of these entities could be visually verified as correct. One in ten PDF documents contained verifiable false entities, mainly as commercial content for online converting services. A third of Word documents contained verifiable false entities such as "standard user" and "test." The larger number of PowerPoint documents with visible creator data present made it possible to validate more entities possible. One in five entities could be verified as either correct or false.

Different AMG approaches to generate "Creator" entities were taken based on visual characteristics. Using the first visible line or the text section with the largest font resulted in correctness rates of between 0% and 3%, varying between the document formats. Extraction efforts based on collection of the content located immediately beneath the correctly identified title resulted in correctness rates of between 4% and 20%.

Word and PowerPoint documents can contain style tags that present the formatting used for specific sections in the document, typically based on template data. No documents contained the style tags "Author" or "Creator". PDF documents also support inclusion of style tags. No PDF documents were found that included the desired tags.

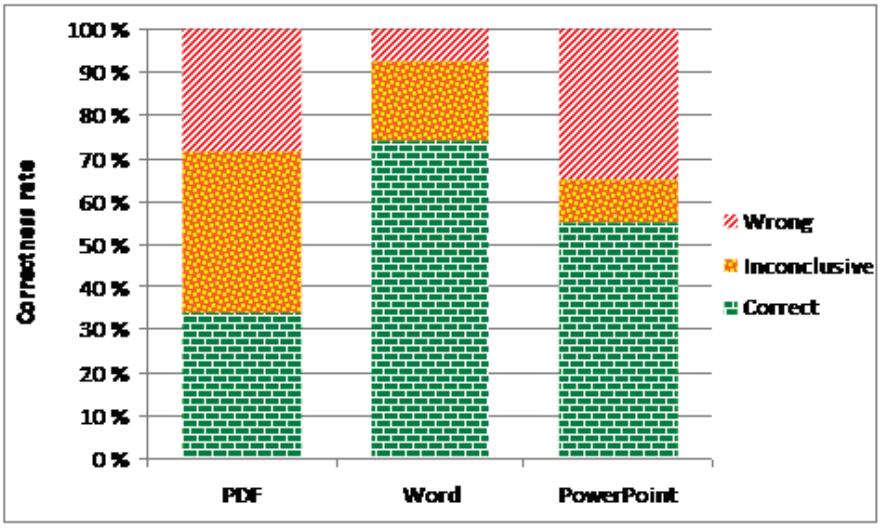


Figure 3: Documents created by LMS publisher

Half of the PowerPoint documents contained “Sub-title” style tags. Two-thirds of all visible creator information was found in this section. These sections were visually formatted in a variety of ways, and contained a range of different data, such as sub-titles, dates, course descriptions and creator information in a multitude of different orders. Creator information was included in 60% of the “Sub-title” sections present. Eight percent of the “Sub-title” sections contained only creator information. The variety in regards to content types and visual formatting makes extraction efforts from this section reliant upon identification of user- and organization names in among the text.

An alternative to generating creator metadata could be the harvesting of context publisher data from the LMS, which could then be used as creator metadata. Such an approach can generate valid entities for individual publishers, although false entities would be generated for groups of authors. The current research compared the LMS’ user name against the embedded metadata and visible characteristics. This approach was able to confirm that three-quarters of Word documents were published by their creators, that PowerPoint documents were more frequently published by others than the document creators, and that there were extensive uncertainties regarding PDF documents. Hence, this approach still produces a great deal of uncertainty and false results.

5.2 Generating “Title” elements

This chapter analyses embedded “Title” entities from common document formats and AMG approaches for generating such entities. This is performed with a special focus on using document code analysis as basis for extraction efforts. The PDF document code showed not to include content relevant for this analysis. These efforts were therefore focused on Word and PowerPoint documents; 200 Word and PowerPoint documents were selected at random. Two corrupted PowerPoint documents could not be analyzed. The remaining documents were losslessly converted to their respective Open XML document formats. The baseline AMG results were generated based on the efforts of related work:

- **File name:** Obtained from the file system (Bird and the Jorum Team, 2006).
- **Embedded metadata:** Harvested from the document (Greenstone, 2007; Google, 2009; Scirus, 2009; Yahoo, 2009; Jenkins and Inman, 2001; Singh et al., 2004).
- **First line:** Extracted from the first visible line of text (Greenstone, 2007).
- **Largest font:** Extracted the text section on the first page based on the largest font size (Giuffrida et al., 2000; Google, 2009).

The results of the baseline efforts were categorized as correct, partly correct, no results and false results:

- **Correct:** The generated entity was identical or nearly identical to the visible title. Small variations, such as spaces that had been removed between words, were accepted.
- **Partly correct:** The generated entity was either partly correct or larger differences were present.

- **No results:** No content was generated by the algorithm. This can be the result of documents without embedded metadata or documents without text-based content.
- **False results:** The generated entity does not result in a representative "Title" element.

The baseline results show that using the content with the largest font generated the most correct entities. The embedded metadata was strongly influenced by being automatically generated the first time the document was stored, and hence was not updated as the document evolved during the creation process. The first line algorithm frequently collected the document header section from page tops.

Table 1: Baseline "Title" results: Word documents

Algorithm	Correct	Partly	No result	False
File name	40%	45%	0%	15%
Embedded	27%	29%	8%	36%
First line	38%	15%	1%	46%
Largest font	69%	8%	1%	22%

Table 2: Baseline "Title" results: PowerPoint documents

Algorithm	Correct	Partly	No result	False
File name	21%	52%	0%	27%
Embedded	28%	10%	0%	62%
First line	37%	34%	2%	28%
Largest font	76%	14%	2%	8%

Open XML documents are zip archives containing standardized, structured content regardless of the document content. There are dedicated XML files for the footer and header sections. As a result, these sections can be avoided entirely. By analyzing the content of the main document XML file of Word and PowerPoint documents, it is possible to analyze the main document content based on facts without the need for visual interpretations e.g. regarding font name and size, placements and section content.

Eight of ten PowerPoint documents contained a “Title” style tagged section. These sections contained nothing but titles, formatted in a variety of different ways. Three percent of Word documents also contained such sections, though two out of three documents used this section for data other than title information.

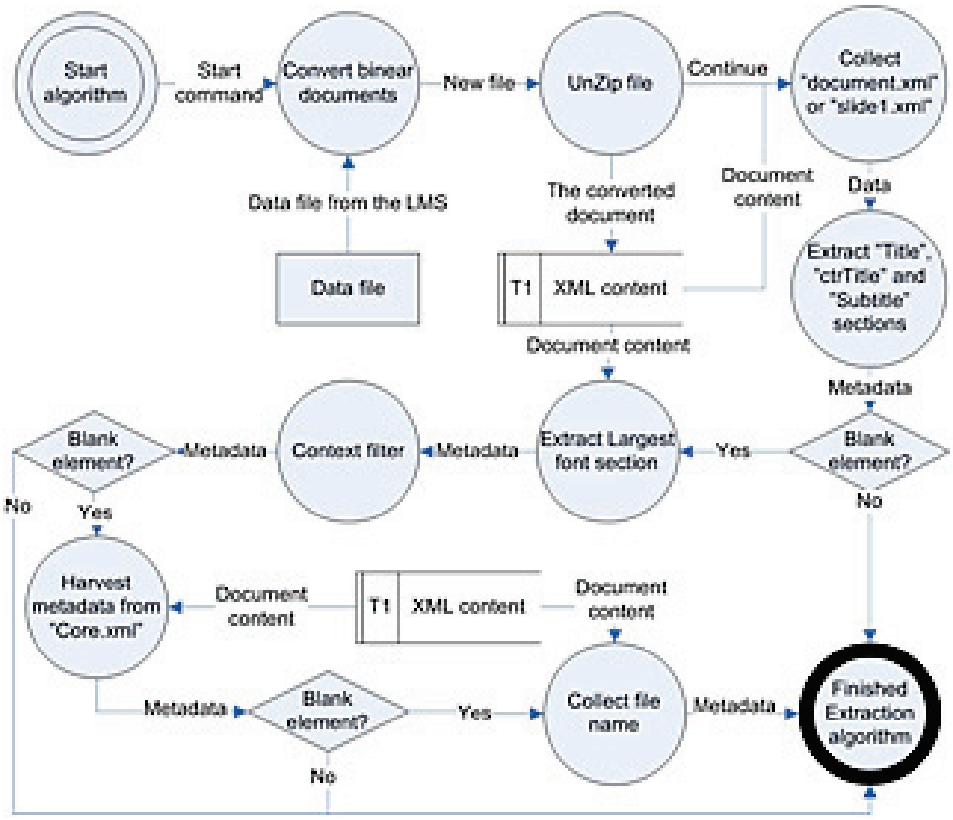


Figure 4: Logical structure of algorithm C

The key property that allows the document code analysis approach to be combined with other AMG methods is that it does not deliver a result when the desired content is not located. This enables it to be combined with other AMG methods. Our research demonstrated this by testing three different document code analysis based algorithms:

1. **Document code exclusively:** Generates “Titles” elements based exclusively on the document code.
2. **Document code and largest font:** Extends algorithm A by evaluating if algorithm A provides an entity. If not, then the content with the largest font section is collected.

3. **Document code, largest font, context filter and alternative data sources:** Extends algorithm B by evaluating if algorithm B provides an entity after performing context data filtering (e.g. course codes and course descriptions). If no entity is generated, then the embedded metadata entity is harvested. If this entity is empty then the file name is used as entity.

The falsely labelled Word document appeared in the algorithm results. As these AMG efforts were constructed to demonstrate the possibilities of using document code analysis, these results have been accepted. The AMG efforts associated with algorithm B focus on documents for which there were no results from algorithm A. This results in a large portion of correct records, though with faults. The inclusion of context data filters in algorithm C reduced the number of false records greatly. One document was given a title based on the file name, since neither the document body nor the embedded metadata contained text-based content. By excluding use of algorithm A for Word documents, the correctness rate would increase by two percentage points, reducing the number of false records by a similar amount.

Algorithm A employed the "Title" style tags that are frequently included in PowerPoint documents. All these sections contained valid titles. The remaining AMG efforts of algorithm B then concentrated on documents that did not have a style formatted title. This resulted in one document being given a false label while three documents received a correct title. Algorithm C gave titles based on the file name to the documents without text-based content. No filtering of content was performed.

Table 3: Basic AMG approach results: Word documents

Algorithm	Correct	Partly	No result	False
A	0%	0%	98%	2%
B	71%	6%	1%	22%
C	91%	6%	0%	3%

Table 4: Basic AMG approach results: PowerPoint documents

Algorithm	Correct	Partly	No result	False
A	85%	0%	15%	0%
B	94%	0%	3%	3%
C	97%	0%	0%	3%

5.3 Generating "General.Language" elements

This chapter presents usage of the existing, automatically generated language tags from common document formats for AMG purposes. The document code can contain tags reflecting the language of the document's intellectual content. This allows for populating the IEEE LOM's "General.Language" element (IEEE LTSC, 2005) and execution of AMG algorithms based on natural language in multi-linguistic environments. Language recognition is automatically performed by applications such as MS Word and MS PowerPoint on document text sections to enable spelling and grammar checks. These section-wise language descriptions are stored as language tags in the documents. Our research documented that language tags are discarded if the document is converted to a PDF. This research is hence focused on Word and PowerPoint documents. One hundred documents were selected at random, resulting in 60 Word and 40 PowerPoint documents. These documents were lossless converted to their native Open XML document format. The analysis was performed on the main document content of Word documents and on the first slide of PowerPoint documents.

All Word documents contained US English language tags, though less than one in ten of the Word documents *used* these tags. Extraction efforts need to be focused on the tags that are in practical use. The extraction effort showed that all text sections were formatted with a single language tag. This allows for using language-specific natural language AMG algorithms on individual sections formatted with a specific language tag. Both single and multi-lingual documents were found.

PowerPoint documents typically contain a limited number of complete sentences for which language recognition can be performed. Hence less data is commonly available to determine the language used in the document. This can result in less accurate language tags than for Word documents. Single language PowerPoint documents were found in Norwegian, US English and British English. One document contained false language tags, when a few Norwegian keywords were included on the first slide of an US English slide show. This illustrates the difficulties of recognizing short language sections. Thirty percent of the PowerPoint documents were correctly labelled as containing multi-lingual intellectual content.

6 EVALUATION

The analysis of Chapters 4 and 5 revealed issues which affect the quality of the metadata which can be automatically generated based on these data sources. This chapter review these issues based on the quality terms of Lindland et al. (1994) and Bruce and Hillmann (2004). Chapter 6.1 presents the embedded metadata. Chapter 6.2 presents the effects of the extraction efforts and Chapter 6.3 summons up the effects of the document creation process.

6.1 Embedded metadata

The documents created in the controlled user environment did not contain embedded metadata. This evaluation of embedded metadata is hence concentrated on stand-alone documents. We observed that embedded metadata was created by applications and users, and inherited from templates and old versions of the documents. None of the document formats analyzed contained meta-metadata. The provenance aspect of the metadata quality was hence very low. The applications could, based on reasoning, be determined to be the author of most of the embedded metadata. Determining the creator of semantic elements was difficult since these elements were free for all parties to use. Standardized entities meant that the metadata creator could be determined in selected document-specific cases.

Each document format has its own approach to embedded metadata. The metadata harvesting efforts therefore needed to be adapted to each document format in order to access, interpret and retrieve the metadata. This reduces the quality of the accessibility of the metadata. It also requires ongoing efforts to adapt the harvesting efforts to new document formats or new versions of the document formats over time.

Our research did not explicitly discover content from the main section of the document (document content) that was syntactic false. However, a few documents were found where the syntactic requirements of the document format were not met. These documents were hence corrupted. These documents became corrupted before or as a part of the transfer process to the LMS.

The security restriction properties of specific PDF documents presented themselves as a hurdle for both harvesting and extraction of metadata. For PDF documents with security restrictions, the semantic quality of the metadata was very low since the metadata are unavailable. Security restrictions also limit the possibilities to extract metadata based on these documents.

Selected PDF documents showed false semantic metadata formatting. This reduces the logical consistency aspect of the metadata quality. However, because these problems were present in a systematic way, error correction can be automatically performed. Semantic issues were discovered regarding characters in the XMP metadata section of PDF documents. This reduces these entities' quality based on accuracy.

This research was able to prove that some of the "Date" related entities were false, which made their quality in terms of accuracy less than optimal. The vast majority of dates could not be verified as correct. A very limited number of documents could be confirmed to have false entities. The semantic quality of the "Date" elements could not be fully verified and hence remains undetermined.

Most of the semantic uncertainties we discovered were in the "Title" element. This element was commonly automatically generated by the applications. The generated entities were of a low semantic quality due to: (1) Timeliness: The metadata could be collected from template data or from earlier versions of the document. This affected the quality in terms of the currency of these elements. (2) Accuracy: The AMG algorithms

used generated entities that do not reflect upon the metadata schema's definition of the element content.

Some applications do not to use the document as data source for generating semantic elements. The quality in accuracy for the "Title" entities was low when compared to the visually presented title. The quality varied between document formats as different applications use the main document's intellectual content in different ways to generate these entities and due to the templates used. The pragmatic quality of these entities from Word and PowerPoint documents was low.

The above issues also affected the "Creator" element. The dataset showed that user creation of manual "Creator" elements was even more limited than for "Title" elements. The entities that are present are often based on applications user names rather than the name of the user. Very few documents had visible creator data, so there were very few documents that could be confirmed as having a valid "Creator" element. The semantic quality of the "Creator" element was thus presumed very low.

None of the document formats analyzed contained metadata on the language of the documents' intellectual content. The metadata quality in terms of completeness was hence very low.

6.2 Extraction efforts

The extraction efforts confirmed that high quality metadata can be generated based on document code analysis, although the "Creator" data were not found as style tags, or was visually present only to a limited extent. There was therefore not enough data for the extraction efforts to perform optimally. This confirmed that extraction efforts, such as Giuffrida et al. (2000), Kawtrakul and Yingsaeree (2005) and Liu et al. (2007) are not able to perform on such a diverse dataset. Using an external data source, such as proposed by Bird and the Jorum Team (2006) and Greenberg et al. (2005), generated higher quality metadata, although still with a large number of errors and much uncertainty.

The content of the style tagged "Title" sections of PowerPoint documents were of very high semantic quality. Such formatting was extensively used by users because this section visually presented in the default PowerPoint templates. We did not observe that Word documents visually promoted document sections. As a direct result, very few used document formatting in accordance with the pre-defined style types. The semantic quality of these formatting tags from Word documents was low. In the LMS' controlled user environment, the "Title" section contained consistently high semantic quality entities, because of no alternative title presentation and since it is mandatory to use. Our analysis confirmed that the document code provided a more accurate approach for extraction efforts, either based on the document code directly, or by combining the document code with other extraction algorithms.

The generation of "General.Language" elements resulted in entities of very high semantic, syntactic and completeness quality for Word and PowerPoint documents. Some uncertainties were found when only short text sections were available.

6.3 Effects of the document creation process

Stand-alone documents provide a user flexibility that is not found in the controlled user environment. This ensures that the users' creative efforts can be used to the fullest to express the intellectual content of the document. The applications used *can* create extensive metadata descriptions and create content with high syntactic and semantic quality. But this creative freedom comes at the expense of the documents' systematic quality properties:

- Templates (or old documents) can contain content (embedded metadata and visible intellectual content) that is false or becomes false when used as the basis for new documents.
- The syntactic quality of the document format cannot be assured due to diverse usage among various applications.
- The user may violate template content and its intended usage.
- Converting original documents can alter, add or remove metadata, formatting data and intellectual content.
- Documents can have security restrictions, which prevent AMG algorithms from accessing the documents' content.

Compared to the controlled user environment, stand-alone documents subjected to AMG efforts require different approaches in treating data sources. The data sources from stand-alone documents can be of a variety of qualities. This makes it essential to learn the characteristics of each document format and its practical usage before AMG efforts are undertaken. Harvesting and extraction efforts based on stand-alone documents are less systematic than those based on documents from the controlled user environment.

7 CONCLUSIONS AND FUTURE WORK

AMG algorithms base their efforts on systematic and consistent properties of the documents at hand in order to generate quality metadata in accordance with pre-defined metadata schema(s). AMG algorithms need to find common structures in which to base their efforts, even if the dataset is not homogenous. Recognition of the most correct and most desirable document properties is the basis for automatic generation of high quality metadata.

This research vastly extends the Stage-of-the-art for using document code analysis for AMG efforts and enabling combination of AMG algorithm types on the same resources, validated against an established framework for defying the resulting data quality. This research has documented that document code analysis can be used to automatically

generate metadata of high quality even though the data source is not homogenous. Common, non-visual document formatting that can be obtained through document code analysis enables the generation of high quality metadata. This code is unique for each document format, although it is shared by all documents of the same document format version. Document code analysis allows for the unique identification of all sub-sections of the documents and enables extraction from each formatted section individually, which in turn allows for the generation of a multitude of different metadata elements. AMG efforts based directly on document code analysis only generate results when the desired content is present, avoids interpretation of the document content and can provide other AMG algorithms document descriptions based on facts. These properties enable efficient combinations of AMG algorithms, allowing different harvesting and extraction algorithms to work together in order to generate the most desired, high quality results.

AMG efforts based on stand-alone documents require an understanding of how the documents are used by the document creators (users), what the user specifies and what is automatically generated based on templates and application specific AMG algorithms. This research has documented that such efforts can generate high quality metadata from stand-alone documents from a non-homogeneous dataset. This research has presented how AMG efforts can be combined in order to generate high quality metadata from a user controlled document creation environment.

The AMG research field is still young and much remains unexplored. At the same time the use of digital documents is increasing dramatically, which offers the potential for extensive research efforts in the years to come. Future work should include (1) Analysis of the impact of the usage environment in which documents are created, and (2) Exploring the possibilities for practical experiments using AMG technologies.

8. REFERENCES

Bird, K. and the Jorum Team. 2006. *Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems*. 31st March 2006, Version 1.0, Final, Signed off by JISC and Intrallect July 2006.

Boguraev, B. and Neff, M. 2000. *Lexical Cohesion, Discourse Segmentation and Document Summarization*. In In RIAO-2000, Content-Based Multimedia Information Access.

Bruce, T.R. and Hillmann, D.I. 2004. *The Continuum of Metadata Quality: Defining, Expressing, Exploiting.* ALA Editions, In Metadata in Practice, D. Hillmann & E Westbrooks, eds., ISSN: 0-8389-0882-9

Cardinaels, K., Meire, M. and Duval, E. 2005. *Automating metadata generation: the simple indexing interface.* In Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, pp.548-556, ISBN:1-59593-046-9

Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. 2007. Automated Template-Based Metadata Extraction Architecture. Proceedings of the ICADL 2007.

Giuffrida, G., Shek, E. C. and Yang, J. 2000. *Knowledge-Based Metadata Extraction from PostScript Files.* In *Digital Libraries*, San Antonio, Tx, 2000 ACM 1-581 13-231-X/00/0006

Google. 2009. *Google*. http://www.google.com

Greenberg J., Spurgin, K., Crystal, A., Cronquist, M. and Wilson, A. 2005. *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project*. UNC School of information and library science.

Greenberg, J. 2004. *Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications.* In Journal of Internet Cataloging, 6(4): 59-82.

Greenstone. 2007. *Source only distribution*. http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected)

IEEE LTSC. 2005. *IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata, WG12: Related Materials*.

It's learning. 2009. *It's learning*. http://www.itslearning.com

Jenkins, C. and Inman, D. 2001. *Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF*. 0-7695-1022-1/01, 2001 IEEE

Kawtrakul A. and Yingsaeree C. 2005. *A Unified Framework for Automatic Metadata Extraction from Electronic Document.* In Proceedings of IADLC2005 (25-26 August 2005), pp. 71-77.

Li, H., Cao, Y., Xu, J., Hu, Y., Li, S. and Meyerzon, D. 2005a. A new approach to intranet search based on information extraction. Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, Pages: 460 – 468, ISBN:1-59593-140-6, ACM New York, NY, USA.

Li, Y., Dorai, C. and Farrell, R. 2005b. *Creating MAGIC: system for generating learning object metadata for instructional content.* Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, pp.367-370, ISBN:1-59593-044-2

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekema, A., McCracken, N.J., Silverstein, J. and Sutton, S.A. 2002. *Automatic metadata generation and evaluation.* Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11–15, Tampere, Finland, ACM Press, New York, pp.401–402.

Lindland, O.I., Sindre, G., Sølvberg, A. 1994. *Understanding Quality in Conceptual Modeling.* In IEEE Software, march 1994, Volume: 11, Issue: 2, pp. 42-49, ISSN: 0740-7459, DOI: 10.1109/52.268955

Liu, Y., Bai, K., Mitra, P, and Giles, C.L. 2007. *TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries.* Proceedings in JCDL'07, June 18–23, 2007, Vancouver, Canada, ACM 978-1-59593-644-8/07/0006

LOMGen. 2006. *LOMGen*. http://www.cs.unb.ca/agentmatcher/LOMGen.html

Meire, M., Ochoa, X. and Duval, E. 2007. *SAmgI: Automatic Metadata Generation v2.0.* In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, pp. 1195-1204, Chesapeake, VA: AACE

Open Archives Initiative. 2004. *Protocol for Metadata Harvesting – v.2.0*. http://www.openarchives.org/OAI/ openarchivesprotocol.html

Scirus. 2009. *Scirus – for scientific information*. http://www.scirus.com

Seymore, K., McCallum, A. and Rosenfeld, R. 1999. *Learning hidden Markov model structure for information extraction.* In Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction, pages 37-42, 1999.

Singh, A., Boley, H. and Bhavsar, V.C. 2004. *LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology*. National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004

Yahoo. 2009. *Yahoo!*. http://www.yahoo.com

P4: Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing?

Published in the Proceedings of IEEE DEST 2010
IEEE Computer Society 2010, pp. 206-221, 2010, ISBN 978-1-4244-5553-9

Lars Fredrik Høimyr Edvardsen

Intelligent Communication AS / Norwegian University of Science and Technology
Department of Computer and Information Science

Oslo / Trondheim, Norway

edvardsen@hotmail.com

Ingeborg Torvik Sølvberg

Norwegian University of Science and Technology
Department of Computer and Information Science

Trondheim, Norway

ingeborg.solvberg@idi.ntnu.no

Abstract — Enabling efficient retrieval and re-usage of digital documents is a major challenge as many documents on the Internet and on Intranets are poorly described with metadata. Manual generation of quality metadata requires skilled human resources, is costly and time-consuming. As a result, metadata related to the documents are too often insufficient or even in-correct. Automatic Metadata Generation (AMG) algorithms could perform similar metadata generation efforts in seconds without the need for human efforts. Submission of conference proceedings commonly includes specifying an extensive range of metadata. Conference proceedings are based on a specific document template with strict usage regulations making them a prime candidate for AMG efforts. This paper evaluates usage of AMG to generate metadata from papers based the MS Word-based IEEE & ACM conference proceedings templates. This enables this research to evaluate if the templates enable efficient AMG efforts, and if the desired paper content is actually retrieved. As authors might not see value in complying with the templates, actual document content can differ from the template specifications.

Keywords-component — algorithms, reliability, experimentation, verification

1. Introduction

Large amounts of scarce human resources are still being used to create similar resources, such as documents [1]; partly because people are not aware of others work, the lack of sharing opportunities, and the inability to retrieve available documents. This situation is, however, dramatically improved the last years due to the large document

collections being made available on the Internet by publishers, organizations and individuals. A major challenge for content management in such collections is the generation of high quality metadata. Manual generation of quality metadata requires skilled human resources. Trained librarians and archivists can assist authors to create and publish metadata, but this is a costly and time-consuming process. Automatic Metadata Generation (AMG) is methods for generating metadata without manual interaction using computer program(s) to interpret the document and possibly the document context. AMG is based on the observation that information that equals the desired metadata, directly or indirectly, may already be contained in the documents or in the context as:

- *Visual descriptions:* By viewing the document through its native application or as a print-out, visual characteristics can be seen, such as the paper format and promotion of specific sections (e.g. some text with larger letters).
- *Technical descriptions:* By analyzing technical information from the document or the system in which the document is stored, other characteristics can be obtained, e.g. file size, file format and dates.
- *Intellectual content descriptions:* By analyzing the user specified textual content of a document, the intellectual content created by the user can be determined, such as the actual letters used to stipulate the document title.
- *Context descriptions:* There is commonly a link between the document which is created and the place in which it is published. E.g. published site and publisher role at that site.

The author of the document has hence directly or indirectly specified the desired content of metadata elements. This can be utilized as AMG strive to avoid excessive manual efforts when similar metadata can be generated automatically based on existing data sources [2, 3, 4, 5, 6, 7, 8]. Previous work have documented that analysis of the document file format, the document code, as seen in Fig. 1, can reveal extensive amounts of information regarding the document without having to use an interpreter application to re-generate the documents' visual appearance [2, 3, 8]. These analyses also show that publishers were not willing to generate more metadata then what was system required, often resulting in low quality metadata in the fields filled out. In another case study, the publishers were not willing to use any time on generating metadata [28].

Paper publishers are commonly forced to generate extensive metadata descriptions. The submission process for a proceedings paper typically consists of filling out dozens of paper descriptions including title, abstract, keywords, author information and a conflict of interest list. Filling out such submission forms can be tedious work consisting of reproducing (copying and pasting) content from the paper and filling out numerous check-boxes and other lists into the submission form.

```
<w:body>
  <w:p w:rsidR="003350B9">
    <w:pPr>
      <w:pStyle w:val="Title" />
    </w:pPr>
    <w:r w:rsidRPr="003350B9">
      <w:rPr>
        <w:lang w:val="en-US" />
      </w:rPr>
      <w:t>Automatic Metadata Generation</w:t>
    </w:r>
  </w:p>
</w:body>
```

Figure 1. Example of simplified document code from an Open XML (MS Office 2007) Word document

In principle, all content in the papers which equals one or more metadata descriptions, could be automatically generated without human intervention using AMG algorithms, saving paper authors for time and efforts while ensuring high metadata quality. Though, this requires two factors:

a) That the AMG algorithm is able to locate the desired paper content section correctly.

b) That the paper content sections are used in a systematic manner.

The document template lay the ground rules for technical and visible formatting, specifying how and where content sections should be present. This research has examined the Microsoft Word document template used for IEEE DEST [31] and the ACM SIG Proceedings Template [29] to see how these templates enable distinction of different content sections. The papers created using these templates can differ from the templates if they are falsely used. 50 actual papers based on the specified templates have been retrieved from the Internet and analyzed to see how these papers comply with their template and ultimately enable correct generation of metadata. This research expects that some of the papers are drafts or reproductions of papers. For convenience, this research used the Conflict of interest list of 2009 for all the papers. This analysis is based on the ACM template only, as sufficient number of IEEE DEST based Word documents were not retrievable.

Chapter 2 presents the state-of-the-art of the field of AMG. Chapter 3 presents the paper templates and how the AMG algorithm is designed to work with these templates. Chapter 4 presents how the algorithm performs on reference papers and on actual papers. Chapter 5 concludes and presents future work.

2. Automatic Metadata Generation

AMG algorithms are sets of rules for processing of data source(s), identification of desired content, and collection and storage of data in accordance with a metadata schema. AMG can be used on all digital resources ("documents"). AMG algorithms base their efforts on systematic and consistent properties of the documents at hand in order to generate quality metadata in accordance with pre-defined metadata schema(s). AMG algorithms need to find common structures in which to base their efforts, even if the dataset is not visually homogenous. AMG algorithms can use the document itself and the context surrounding the document as data sources. Collecting embedded metadata is known as metadata harvesting [5, 9]. The process by which AMG algorithms create metadata that previously has not existed is known as metadata extraction [10, 11]. AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that is incorrect for the description of a document. Four different approaches are used to generate metadata, as illustrated in Fig. 2.

Harvesting of embedded metadata: This approach uses the embedded metadata created by applications or by the user and stored as part of the document [11, 12, 13, 14, 15, 16]. This research will extend this approach by providing the Intranet used for data storage with educational, contextual information which can be retrieved and used for AMG purposes directly or indirectly. Harvesting of embedded metadata is vulnerable to generating false metadata if the embedded metadata is incorrect.

Extraction based on visual characteristics: This approach uses a content presentation application to create a visual representation of the document before executing rules to extract content based on the visual characteristics of the document [17, 18, 19, 20, 21]. This approach is vulnerable to generating false metadata if the documents do not share the visible characteristics with which the algorithm has been developed to perform. Hence, such algorithms only perform as desired on pre-known document types and not on document is general.

Extraction of metadata based on natural language: This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed [22, 23, 24, 25, 26, 27]. Such algorithms commonly include collection of unique words and comparisons of the document vocabulary against reference ontology for generating keywords, descriptions and subject classification. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages, document sections in different languages or contains header or footer fields since the text from these fields are presented on every page and hence occur frequently.

Extraction based on document code analysis: This is a new approach presented by this project. This approach uses analysis of the code of the document directly without the need for additional content presentation applications to interpret the document content [2, 3, 8]. This enables full and direct access to the entire document's content. This

includes template identification, template content identification and formatting characteristics regardless of visual characteristics, and the language of the intellectual content. It also enables effective usage of other AMG approaches when the document code does not provide the desired data. This approach requires that the document code is understandable for the AMG algorithm and that the end-user does not misuse the document template.

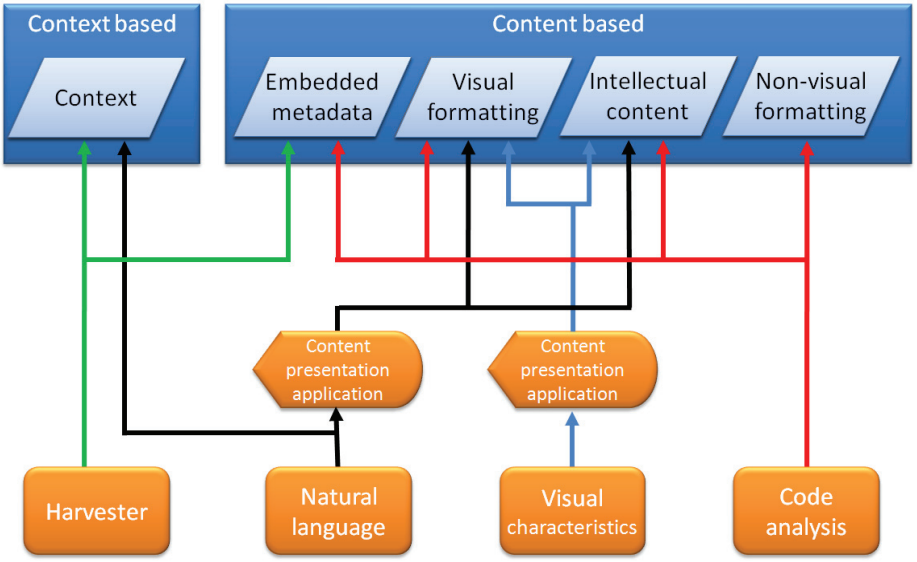


Figure 2. AMG analysis algorithms and the data sources which they use

This project has been using the characteristics of these state-of-the-art AMG algorithms in order to gain a collection of baseline results in which the efforts of this project can be compared against.

3. Automatically Generating Paper Metadata

3.1 The Paper Templates

The IEEE DEST uses a Word template to ensure correct document formatting. The ACM SIG uses Latex and Word templates. The Latex template contain specific and named sections to identify each author, address fields, the title, general terms etc. using special formatting tags and non-visual descriptions. This makes it a prime candidate for

AMG efforts. Though, the Latex format is not distribution friendly due to e.g. multiple files needed to recreate a single document. Hence, Latex documents in general are commonly converted to a PS or PDF-file before publication. In this conversion process, the template's original formatting tags and non-visual descriptions are discarded, loosing information usable for AMG algorithms. The embedded document metadata is commonly corrupted [3].

Word documents do not contain full visible information as PS or PDF-files. Though, they end up being more informative for AMG efforts as their original visible and non-visible formatting consists within the published file. The analyzed Word templates promote virtually the same document sections, though with slightly different visible presentation. The IEEE DEST template does not use formatting tags to identify individual sections. The ACM SIG Proceedings Template contains non-visual style tags labelled for formatting the title, abstract and e-mail address sections. No visible or non-visible section tagging is used for other sections, such as general terms or keywords.

3.2 Creating a Paper registration AMG algorithm

All tagged sections in Word documents are re-locatable in the Document Code of the document file. The content of the tagged sections should be retrievable regardless of the document's visible appearance. By lossless converting the Word-documents into the OpenXML Office Word 2007 format, the complete content of the original file's Document Code can be analyzed [2, 3, 8]. This research created an AMG algorithm to test if the content of the tagged sections actually were retrievable, and usable as a substitute for manual metadata registration efforts while ensuring high metadata quality. If no formatting tags are located with textual content, then the AMG algorithm is unable to generate metadata for this element using this method.

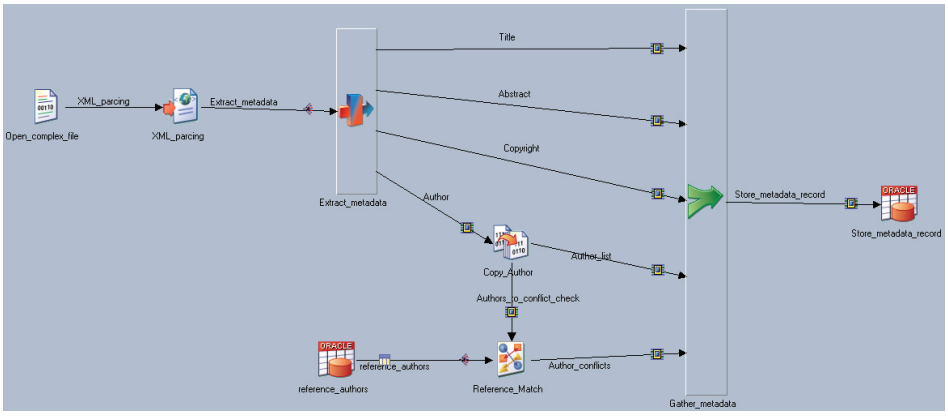


Figure 3. Simplified execution sequence

The formatting tags were additionally used as key section indicators to guide extraction based on visible characteristics when the desired content were not tagged. E.g. the author name(s) should be on the line directly under the title line. This way, AMG efforts based on the Document Code were executed first, and then efforts based on visible characteristics, in line with previous experiences [2, 3, 8]. A drawback of using AMG algorithms based on visible characteristics is that entities are generated as long as content is located. This is a known source to false entities [2, 3, 8]. IBM QualityStage software [30] was used to execute the sequence three main sections of the execution sequence based on the execution sequence structure presented in Fig. 3:

1) **Submission:** The paper is submitted as a Word document to the framework and made available for the AMG algorithm.

2) **AMG pre-processing:** The Word-document is converted into OpenXML. Desired document sections are identified based on formatting tags and their content retrieved.

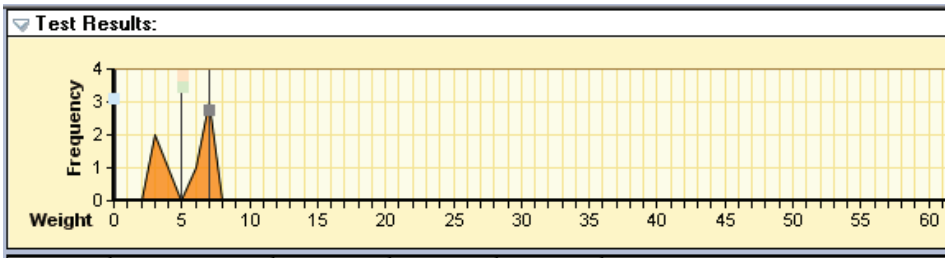
3) **AMG post-processing:** The available data are used to generate new metadata in order to populate such as the conflict-of-interest list.

The framework was created to populate the following elements: “Title”, “Author”, “Affiliation”, “E-mail”, “Abstract”, “Categories”, “General terms”, “Keywords”, “Copy right” and “Conflict of interest”.

In paper submission forms there is commonly a distinction between “First name(s)” and “Sur name”. Such a distinction is not present in any of the paper templates, using the single field “Author name” instead. Due to different cultural and national standards, it is not always valid to claim that the last name always is the last “word” in the “Author” entities. In this research, the AMG algorithm does not attempt to distinguish between First- and Sur name.

Selected metadata elements need extra processing in order to gain the desired entities. Observations showed that multiple authors and e-mail addresses are commonly included on a single line. Due to this, the AMG algorithm has included logics to separate such lists based on common entity separators, such as commas and the “&” sign.

Extraction of entities for the “Conflict of interest” lists also required additional efforts. A reference list of people was created. By using a statistically weighted “Multiple uncertainty” comparison algorithm, the author names were compared against each of the pre-registered persons in the reference list. The algorithm generated a weight “score” for each comparison, giving statistically more similar and unique names higher scores. A weight score threshold was enforced to ensure that only matching results with results over a specified target were accepted as a match. Fig. 4 show the comparison results from a test paper. All matching results with weight below five were specified to be discarded. All matching results over the threshold were retrieved and used to populate the Conflict of interest list. A complete execution sequence finishes in a matter of seconds per document.



SetID	Record Type	Weight	DataID	RefID	Author And RefAuthor
5	MA	7,00	5	28	Fox Ed
5	MB	7,00	5	28	Edward Fox
6	MA	7,00	6	78	Ingeborg T. Solvberg
6	MB	7,00	6	78	Ingeborg T. Solvberg
7	MA	7,00	7	92	Ian Witten
7	MB	7,00	7	92	Ian Witten
2	MA	6,00	2	78	Ingeborg Torvik Sølvberg
2	MB	6,00	2	78	Ingeborg T. Solvberg
1	MA	4,00	1	48	Lars Fredrik Høimyr Edvardsen
1	MB	4,00	1	48	Alberto H. F. Laender
3	MA	3,00	3	78	Trond Aalberg
3	MB	3,00	3	78	Ingeborg T. Solvberg
4	MA	3,00	4	78	Hallvard Trætteberg
4	MB	3,00	4	78	Ingeborg T. Solvberg

Figure 4. Setting threshold for reference matching results

4. Execution Results

The developed AMG algorithm was created to perform strictly in accordance with the specified templates. Hence, if the templates were used as specified, then all entities should be correctly extracted.

4.1 Reference Papers

To test the AMG algorithm, a selection of reference papers was created using the IEEE DEST and ACM templates. These reference papers were created in accordance to the template specifications. By following these templates, formatting tags should be created automatically, and the visible presentation should be in line with the template specifications. In addition, a selection of reference papers was created which deliberately broke with the template specifications. Such issues included placing the title on page number two instead of on page one.

The value of using the Document Code as basis for AMG efforts soon became apparent: All the template specified tagged sections were correctly identified within the reference documents, and their content correctly extracted. Even the title on page two were

correctly identified and extracted, as this was the first tagged title section within the document. The AMG efforts based on visible characteristics also generated favorable results. Though, there were issues when the visible formatting did not comply with the template specifications. This since the desired document content was not located at the expected location. E.g. the specification "author name(s) should be on the line directly under the title line" gave false results when the title was falsely located on page nr two.

4.2 Actual usage of the templates

Analysis of actual, non-reference papers revealed that paper authors do not follow the templates strictly. Most of the papers kept to the visible formatting of the templates. Though, there were exceptions where sections were moved. In addition, there were a number of papers missing specific sections and papers that used other formatting than the template specified, as summed up in Fig. 5.

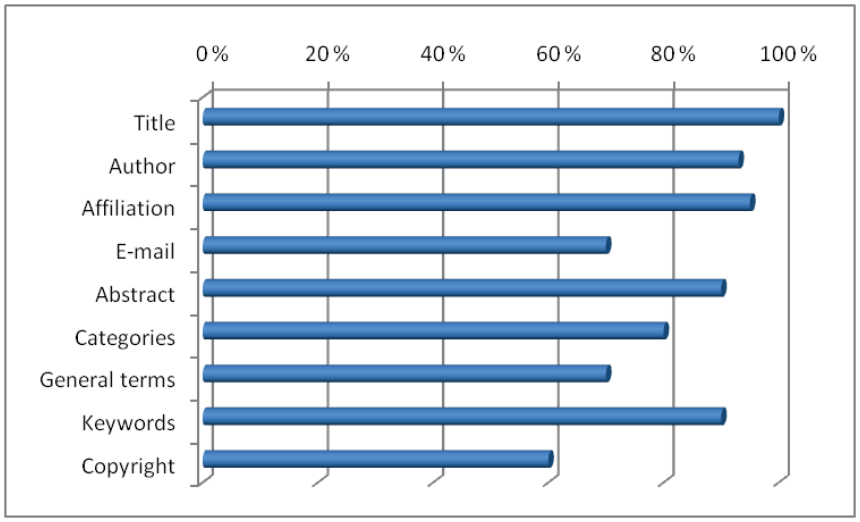


Figure 5. Presence of correctly formatted content in the dataset

The template states that the title should be located on the top of the first page. Selected papers included a comment before the title. As these comments were not tag formatted as a specific content type and the Title element were generated using the Document Code, the inclusion of these did not affect the performance of the AMG algorithm.

All the papers had at least one author registered, though these could be falsely formatted as e-mail addresses or affiliation. One in ten papers contained Author or Affiliation entities that did not comply with the template specification. E-mail addresses were

commonly formatted as “Affiliation” or missing. One in five papers contained falsely formatted e-mail addresses. The falsely formatted content did have a negative effect on the AMG algorithm performance. A result of this was that too many “authors” were registered, as false content were included as an author. There were therefore extra many “authors” which were compared against the reference list of people in order to generate the “Conflict of Interest” list. The result of these comparisons were without faults even though there were some misspellings, shortenings and different word ordering compared to the reference lists.

The Copyright notification message was either falsely visually formatted (commonly included in the first paragraph) or missing from 40% of the documents. A number of authors also skipped filling out the General terms sections. Neither of these sections was format tagged in the template. Hence, AMG efforts based on visible characteristics were executed for these sections. The lack of desired sections resulted in false content being retrieved.

One of the retrieved papers were a work-in-progress document where there were a number of so-called “reviews” present with content registered as added and deleted from an earlier version of the paper. When a Word document contains review data, this is also affected in the Document Code by inclusion of the original document plus all content that has been added and deleted, marked with their own set of formatting tags. The AMG algorithm was not created for handling review issues. As such, no logics were included in the algorithm to process review data. The content of these sections were therefore not processed, resulting in false or incomplete metadata entities.

5. Conclusion and Future Work

Enabling efficient retrieval and re-usage of digital documents is a major challenge as many documents are poorly described with metadata. AMG can play a key role in describing documents with metadata as a substitute to manual efforts, enabling more efficient retrieval and re-usage services. By “labeling” specific content with meta formatting tags, AMG efforts based on the Document Code can locate, retrieved and used as metadata or as basis for generating metadata, regardless of the visible characteristics of the document. Basing AMG efforts on visual appearance is possible, though is more vulnerable to documents which differ from the template specified appearance. Document templates are not just about looks – They are also a specification of the document functionality. The document template specify the common rules for which meta formatting tags that should be present in the document and how these tags should be used by AMG algorithms.

Analysis of the IEEE DEST and ACM paper templates in the MS Word document format revealed that Meta tagging of paper specific content is not commonplace. Only a selected few sections are tagged, even though there are a significant number of sections which could have been tagged in order to enable extraction of detailed metadata descriptions.

Paper authors have also yet to discover the benefits of using the formatting specified in the templates; even with strictly specified document specifications, authors use the templates in a number of different ways.

This research has documented that an AMG algorithm can identify, extract and create metadata which equals the author specified content in a matter of seconds. This could greatly reduce the complexity, knowledge and time issues currently present regarding registration of papers. Though, to enable the potential of AMG, the templates that are the basis for the published documents need to be updated with detailed Meta tag formatting for uniquely identification of desired document content.

With updated templates and an AMG algorithm present in the publishing tool (web site), authors can experience benefits of complying with the specified template as registration form sections could be populated with extracted metadata automatically. This way, authors that follow the template is "awarded" with a faster and easier publishing process. Future work should look into this potential scenario in order to evaluate if end user see a value in AMG and complying with the template specification. Future work should also include creation of an operational AMG system for submitting documents in order to gain more end user experiences.

6. REFERENCES

[1] M. Hämäläinen, B.A. Whinston and S. Vishik, "Electronic markets for learning: education brokerages on the Internet", Communications of the ACM archive, Volume 39, Issue 6 (June 1996), pp. 51 – 58, ISSN:0001-0782, ACM (1996)

[2] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg and H. Trætteberg, "Using the structural content of documents to automatically generate quality metadata", Proc. of Webist 2009, March 23-26, pp. 354-363, ISBN: 978-989-8111-83-8, ACM

[3] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg and H. Trætteberg, "Automatically generating high quality metadata by analyzing the document code of common file types", Proc. of JCDL 2009, June 15-19, ACM

[4] K. Cardinaels, M. Meire and E. Duval, "Automating metadata generation: the simple indexing interface", Proc. of the 14th international conference on World Wide Web, Chiba, Japan, 2005, pp.548-556, ISBN: 1-59593-046-9

[5] J. Greenberg, "Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications", Journal of Internet Cataloging (2004), 6(4): 59-82 M. Meire, X. Ochoa and E. Duval, "SAmgI: Automatic Meta-data Generation v2.0", Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, pp. 1195-1204, Chesapeake, VA: AACE

[6] E. Duval and W. Hodgins, "Making metadata go away: Hiding everything but the benefits". Keynote address at DC-2004, Shanghai, China, 2004

[7] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg and H. Trætteberg, "Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects". IEEE DEST 2009

[8] L.F.H. Edvardsen and I.T. Sølvberg, "Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment", Proc.of WEBIST 2007, March 3-6, ISBN 978-972-8865-77-1, pp. 427-432, Springer

[9] IEEE LTSC, "IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata. WG12: Related Materials", 2005, http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

[10] ADL, "Sharable Content Object Reference Model (SCORM) 2004 3rd Edition Documentation Suite", 2006, http://www.adlnet.gov/downloads/AuthNotReqd.aspx?FileName=SCORM.2004.3ED.DocSuite.zip&ID=237

[11] Open Archives Initiative, "2004 Protocol for Metadata Harvesting – v.2.0", 2004, http://www.openarchives.org/OAI/-openarchivesprotocol.html

[12] K. Seymore, A. McCallum and R. Rosenfeld, "Learning hidden Markov model structure for information extraction", Proc. of AAAI 99 Workshop on Machine Learning for Information Ex-traction, pp. 37-42, 1999.

[13] Greenstone, Source only distribution. http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected), 2007

[14] K. Bird and the Jorum Team, "Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems. 31st March 2006, Version 1.0, Final", JISC and Intrallect July 2006.

[15] Google, "Google", 2010, http://www.google.com

[16] Scirus, "Scirus – for scientific information", 2010, http://www.scirus.com

[17] Yahoo, "Yahoo!", 2009, http://www.yahoo.com

[18] A. Singh, H. Boley and V.C. Bhavsar, "LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology", National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004

[19] G. Giuffrida, E.C. Shek, and J. Yang, "Knowledge-Based Meta-data Extraction from PostScript Files", Digital Libraries, San Antonio, Tx, 2000, ACM 1-581 13-231-X/00/0006

[20] A. Kawtrakul and C. Yingsaeree, "A Unified Framework for Automatic Metadata Extraction from Electronic Document", Proc. of IADLC2005, 25-26 August 2005, pp. 71-77.

[21] P. Flynn, L. Zhou, K. Maly, S. Zeil and M. Zubair, "Automated Template-Based Metadata Extraction Architecture". ICADL, 2007.

[22] H. Li, Y. Cao, J. Xu, Y. Hu, S. Li and D. Meyerzon, "A new approach to intranet search based on information extraction", Proc. of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, pp. 460-468, ISBN:1-59593-140-6, 2005, ACM New York, NY, USA.

[23] Y. Liu, K. Bai, P. Mitra and C.L. Giles, "TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries", JCDL'07, June 18–23, 2007, Vancouver, Canada, ACM 978-1-59593-644-8/07/0006

[24] B. Boguraev and M. Neff, "Lexical Cohesion, Discourse Segmentation and Document Summarization", 2000, RIAO.

[25] LOMGen, "LOMGen", 2006. http://www.cs.unb.ca/agentmatcher/LOMGen.html

[26] J. Greenberg, K. Spurgin, A. Crystal, M. Cronquist and A. Wilson, "Final Report for the AMeGA (Automatic Metadata Generation Applications) Project". 2005, UNC School of information and library science.

[27] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg, H. Trætteberg, "Using Document Code to automatically generate high quality metadata: An Auditing case study", In review, 2010.

[28] ACM inc., "ACM SIG Proceedings Template", 2010, http://www.acm.org/sigs/publications/proceedings-templates

[29] IBM, "IBM – IBM InfoSphere QualityStage – InfoSphere QualityStage – Software", 2010, http://www-01.ibm.com/software/data/infosphere/qualitystage/

[30] IEEE DEST, "IEEE DEST Word document template", 2010, http://dest2010.debii.curtin.edu.au/images/stories/word/paper_formatting_guidelines_and_sample.doc

P5: Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects

Published in the Proceedings of IEEE DEST 2009
June 1-3, 2009, INSPEC, ISBN: 978-1-4244-2345-3, 10.1109/DEST.2009.5276729

Lars Fredrik Høimyr Edvardsen [1, 2], Ingeborg Torvik Sølvberg [2], Trond Aalberg [2] and Hallvard Trætteberg [2]

[1] Intelligent Communication AS, Akersgaten 49, Oslo, Norway,
e-mail : lars.edvardsen@intelcom.no

[2] Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway,
e-mail: ingeborg.solvberg ; trond.aalberg ; hal @idi.ntnu.no

Abstract — The Internet is packed with resources which can be used for educational purposes, referred to as Learning Objects (LOs). Locating *the* LO which is best suited for your educational purposes can be extremely challenging. This since the context surrounding the LO in regards to intended user group, educational level etc. are not included in the resource. The SCORM standard has changed this by including contextual metadata as part of the resource. However, SCORM LOs are scarcely created, much as a result of high knowledge and time requirements needed for creating the necessary metadata. This research has been using Automatic Metadata Generation tools to assist in the metadata creation process, enabling LOs to share common contextual metadata while receiving additional high quality LO specific metadata without the need for manual metadata creation efforts.

Index Terms — Algorithms, Reliability, Experimentation, Verification.

1. Introduction

Enormous amounts of scarce human resources are used to create similar resources, such as documents, repeatedly [1]; This comes about by people not being aware of others work, lack of sharing opportunities, and inability to retrieve available work. Computer aided sharing of digital resources has enabled sharing and reuse of resources across intranets and other large scale document storage and retrieval services. A major challenge for content management in such storage facilities is the generation of high quality metadata; Data which describes the available resources. Manual generation of metadata is human resource demanding and is often viewed by collection managers and

resource ("document") authors as inefficient use of their time. As a result, the Learning ecosystem of digital resources runs inefficiently due to limited sharing and reuse of experiences and the resources which reflects these experiences. There is a desire for other ways to create the needed metadata. Automatic Metadata Generation (AMG) is methods for generating metadata without manual interaction using computer program(s) to interpret the document and possibly the document context. AMG is based on the observation that information that equals the desired metadata often already is contained in the documents, such as:

- *Visual and technical descriptions*: E.g. formatting information and the number of visual pages.
- *Intellectual content descriptions*: User specified textual content. E.g. the document title and author.
- *Context descriptions*: E.g. published site and publisher role at that site.
- The resource author has hence directly or indirectly specified the desired content of many metadata elements. Based on this, why should we manually reproduce something which is already available? AMG strive to avoid excessive manual efforts when similar metadata can be generated automatically based on existing data sources [2, 3, 4, 5, 6, 7].

Each year thousands of LOs are published at the Learning Management System (LMS) at the Norwegian University of Science and Technology (NTNU). Due to the extensive range of subjects and educational levels taught at the University, every educational context is regarded as unique. Contextual LO descriptions are hence critical for widespread reuse. The NTNU LMS, named It's:learning [8], has native support for uploading SCORM LOs. Still, only a handful of LOs is submitted as SCORM LOs to the LMS, none of these created at the University. In order to promote local and global sharing and reuse of existing resources published at NTNU, this research has desired to automatically generate SCORM LOs by combining available LOs with automatically generated metadata labelled in accordance with the IEEE LOM metadata schema standard [9, 10]. The IEEE LOM is extensive, enabling rich and detailed resource descriptions, while being backwards compatible with the more general Dublin Core schema [11]. The IEEE LOM is also a prime example of a metadata schema which can be difficult and time consuming to fill out by end users. Through analysis of content collected from the NTNU LMS, this research has developed a framework of ways in which to automatically generate high quality metadata without being a burden on the publishers and document authors.

In this paper the focus is on retrieval of an existing, published, stand-alone LO and automatically generating metadata from various data sources in order to generate a finished SCORM LO. This includes Harvesting of contextual metadata from the LMS, Extraction of entities from the LOs themselves and combining AMG approaches. As a result, a complete set of IEEE LOM entities are generated, including semantic and technical metadata elements. This research has been implemented, tested and evaluated. It extends previous work [2, 3] by including contextual metadata and generation of SCORM LOs.

Chapter 2 presents the IEEE LOM and SCORM standards, Chapter 3 presents Automatic Metadata Generation, Chapter 4 presents how an educational context can be built using available resources, Chapter 5 presents Harvesting possibilities, Chapter 6 presents Extraction possibilities, Chapter 7 presents generation of SCORM LOs, while Chapter 8 concludes and presents future work.

2. The IEEE LOM and SCORM standards

The IEEE Learning Object Metadata (LOM) is an extensive and complicated metadata schema. There are 45 basic elements, although a number of these elements have sub-elements that make them suitable for multiple usage areas. The IEEE LOM was developed by [13, 14, 15] and has the support and potential to be the standard for LO metadata exchange in the years to come. A variety of local, national versions of the IEEE LOM has been or is under development, including the UK LOM [16], NORLOM [17], SWELOM [18] and CanLom [19]. Metadata is classified into 9 categories with sub-elements: (1) General, (2) Life Cycle, (3) Meta-metadata, (4) Technical, (5) Educational, (6) Rights, (7) Relation, (8) Annotation, (9) Classification. The most major drawback of using this schema is that it requires extensive user pre-training and that it is labor intensive to generate metadata records [2, 3, 20]. IEEE LOM metadata records are known to take more than one hour to manually label [12]. This cripples active use of the schema.

The Sharable Content Object Reference Model (SCORM) [21] specification is a collection of standards and specifications for web-based e-learning of the Advanced Distributed Learning (ADL) Initiative [22]. SCORM uses the IEEE LOM schema for describing the LOs' metadata. In its basic form, a SCORM LO is a Zip-compressed file which contains one or more LOs and a XML-file containing the LMS execution data and IEEE LOM metadata. The metadata should provide the end user with all the information which they need to use the LOs in accordance with the LO authors intentions. Hence, SCORM LOs are extremely user friendly in terms of providing the educational audience extensive information regarding the LO's usage and educational context. However, *manual creation* of these is not developer friendly. As a result, there is a lack of SCORM object even though the LOs are available.

3. Automatic Metadata Generation

AMG algorithms are sets of rules for processing of data source(s), identification of desired content, and collection and storage of data in accordance with a metadata schema. AMG can be used on all digital resources ("documents"), including LOs. AMG algorithms can use the document itself and the context surrounding the document as data sources. Collecting embedded metadata is known as metadata harvesting [5, 23].

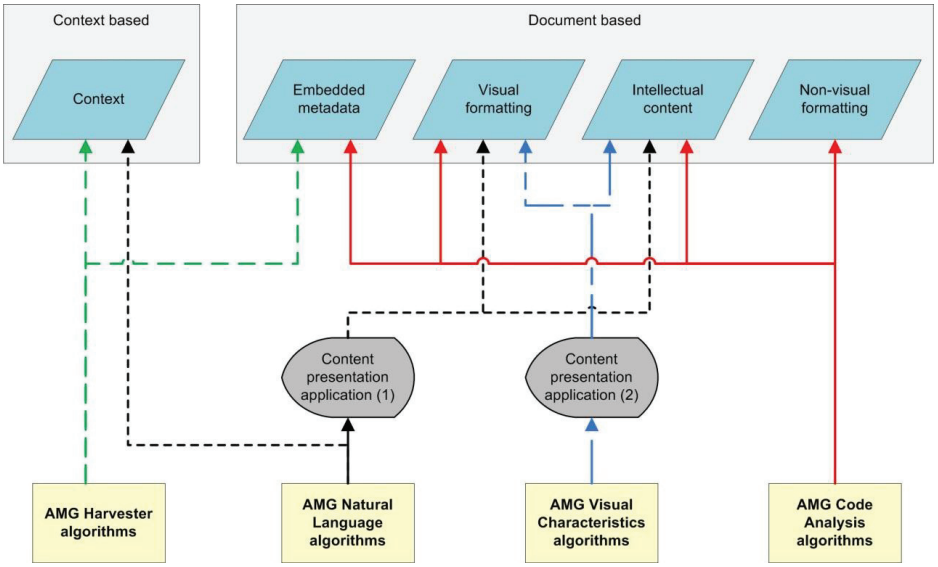


Fig. 1. AMG content analysis algorithms and the data sources which they use.

The process by which AMG algorithms create metadata that previously has not existed is known as metadata extraction [24, 25]. AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that is incorrect for the description of a resource. Four different approaches are used to generate metadata, as presented in "Fig. 1":

- *Harvesting of embedded metadata*: This approach uses the embedded metadata created by applications or by the user and stored as part of the resource [25, 26, 27, 28, 29, 30]. This research will extend this approach by providing the LMS with educational, contextual information which can be retrieved and used for AMG purposes directly or indirectly. Harvesting of embedded metadata is vulnerable to generating false metadata if the embedded metadata is incorrect.
- *Extraction based on visual characteristics*: This approach uses a content presentation application to create a visual representation of the document before executing rules to extract content based on the visual characteristics of the document [31, 32, 33, 34, 35]. This approach is vulnerable to generating false metadata if the documents do not share the visible characteristics with which the algorithm has been developed to perform. Hence, such algorithms only perform as desired on pre-known document types and not on document is general.
- *Extraction of metadata based on natural language*: This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed [36, 37, 38, 39, 40, 41]. Such algorithms commonly include collection

of unique words and comparisons of the document vocabulary against reference ontology for generating keywords, descriptions and subject classification. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages, document sections in different languages or contains header or footer fields since the text from these fields are presented on every page hence occur frequently.

- *Extraction based on document code analysis*: This approach uses analysis of the code of the document directly without the need for additional content presentation applications to interpret the document content [2, 3]. This enables full and direct access to the entire document's content. This includes template identification, template content identification and formatting characteristics regardless of visual characteristics, and the language of the intellectual content. It also enables effective usage of other AMG approaches when the document code does not provide the desired data. This approach requires that the document code is understandable for the AMG algorithm and that the end-user does not misuse the document template.

4. Building an Educational Context of Harvestable Metadata

The NTNU LMS consists of course-specific sections where lecturers and students can freely publish digital LOs. Publishing at each course section is only allowed for users who are logged in with University given user identification and has been given a course specific role, such as "student" or "lecturer". Each course section is named after the course name, e.g. "IT3803 Digital Library".

The University uses a Course catalogue to inform its students of available courses, their subjects and student pre-knowledge requirements etc. This is information which this research has used to enrich the LMS' course information. Here, descriptions have been given on different levels, enabling related courses to the share common descriptions. E.g. all courses inherit the University-specific entities as presented in "Tab. 1" before being populated with course-specific entities as in "Tab. 2". Such reference tables were also used to promote default entities, such as in "Tab. 3".

Tab. 1. University context entities

IEEE LOM elements	Entity
5.6 Context	Higher education
5.7 Typical age range	18-
9.2.2 Taxon	"University", "NTNU"

Tab. 2. Course-specific entities

IEEE LOM elements	Entity
5.5 Intended End User Role	Learner
5.8 Difficulty	Very difficult
5.11 Language	NO
9.2.2 Taxon	"Institute", "IDI"
9.2.2 Taxon	"Course", "IT3803 Digital Library"

Tab. 3. Default entities

IEEE LOM elements	Entity
3.3 Metadata schema	LOMv1.0
6.1 Cost	No
6.2 Copyright and other restrictions	Yes

Tab. 4. University context entities

IEEE LOM elements	Entity
1.5 Keyword	("en", "Digital Library"), ("en", "DL"), ("en", "Metadata"), ("en", "FRBR"), ("en", "Dublin Core"), ("en", "Interoperability")

The "5.8 Difficulty" element entity indicates that this is an advanced course, while the "5.5 Intended End User Role" and "5.11 Language" present that the course is intended for students ("Learners") who speak Norwegian. This hence, is information which describes the context of the LOs, not the LOs themselves. Default entities for schema, cost and copyright specifies that the metadata is generated in accordance with the IEEE LOM v1, that the LO can be used without a fee, though this is a copyrighted LO.

In addition, the Course catalogue contains a short summary of the course and the subjects which it addresses. This research retrieved this course summary and used a Natural Language algorithm to filter out unwanted, common words. This was done in order to generate course-specific keywords, as presented in "Tab. 4". The "en" in the beginning of each keyword indicates that the keyword is in English.

5. Harvesting Metadata

A) Educational Context Data

The course specific contextual entities are made available for the AMG Harvesting algorithms when generating the LO specific entities. The contextual entities are used as a starting-point for LO specific AMG efforts, building a context around the LO. Most of the contextual entities are valid for all LOs published at the specific source. This is however not the case in regards to the "1.5 Keyword" element, as it frequently contains a range of keywords which exceeds the individual LO. This will be further presented in chapter 5.

Due to the log-in features of the LMS, publishing information is stored for all LOs. This is information which is harvested for generation of metadata. "Tab.5" presents such publishing metadata.

Tab. 5. University context entities

IEEE LOM elements	Entity
2.2 Status	Final
2.3 Contribute	"Publisher", "Lars F. H. Edvardsen", "2009-01-31 12:05"
8. Annotation	"Lars F. H. Edvardsen", "2009-01-31 12:05", "Published by course lecturer"

The LMS only supports a single version of a LO, hence each LO can be given a default

"2.2 Status" as "Final". The "2.3 Contribute" element present the contributor and when the contribution was registered. The "8. Annotation" field is used to promote the role of the user in regards to the publication.

B) LO specific Data

Previous work has shown that harvesting of semantic elements from common document formats such as Adobe PDF, MS Word (DOC) and MS PowerPoint (PPT) files commonly result in metadata entities of low quality [3]. This is due to user applications which commonly do not actively use the available elements, used elements are not updated as documents are update and false use due to usage of false content for generating the entity or inclusion of commercial content. This included elements for "Creator", "Title" and "Date". This was also the case for technical elements which were dependent upon correct updating as the document is re-stored, e.g. statistical entities such as "Number of pages" and "Number of words". The previous work did however confirm that technical entities collectable through the publisher file- or database system and static elements from files were of high quality. As a result of this, the elements in "Tab. 6" can be generated.

The "4.1 Format" reflects upon the file format of the LO, while "4.4.3 Minimum version" present the requirement of using an application compatible with version 1.5 of the PDF-format. "4.2 Size" specifies the number of bytes in the LO, while "4.3 Location" presents where the physical LO can be located.

Tab. 6. University context entities

IEEE LOM elements	Entity
4.1 Format	application/pdf
4.2 Size	432304
4.3 Location	http:\\www.ntnu.no\storage…
4.4.1.3 Minimum version	1.5

6. Extracting Metadata from the LO

Analyzing the document code enables access to both the visible and non-visible content of a document [2, 3]. This enables access directly to the intellectual content of the document without contamination caused by the interpretation of the content presented

by content presentation applications. Commonly used LO creation applications, such as Microsoft Word and PowerPoint, include extensive amounts of non-visual information which can be used for AMG purposes. E.g. the user specified title and tables are formatted in a particular way. These sections can be the focus of AMG efforts, regardless of the visual characteristics of the LO. As a result, the "1.2 Title" element can be populated with the user specified title, and the "1.4 Description" element can be populated with the table content.

Modern document creation applications use a range of tools to aid in the document creation process. E.g. spelling- and grammar checks are common to avoid false or bad spelling. Here different algorithms examine the user specified intellectual content in order to determine the language of this content. The results of using such aids can be stored as part of the resulting document as "language tags" [2, 3]. These language tags can be retrieved in order to populate the "1.3 Language" elements which represents the intellectual content of the LO.

In previous work this research used a combination of AMG approaches to generate Title entities [2]. This was achieved by using document Code analysis as basis for Extraction efforts. If no desired content were located, e.g. there was no Style formatted "Title" section, then the document code made the basis for other AMG efforts primarily based on visual characteristics. For other elements, other combinations of AMG algorithms were used. E.g. keywords were generated using a document Code analysis to generate the suitable data foundation for a Natural Language algorithm to execute by:

- Removing non-desired document sections, such as headers and footers
- Directing The Natural Language algorithm to the document sections which contained intellectual content in the same language as the Natural Language algorithm was designed to be executed against.

AMG algorithms based on Natural Language commonly use frequency of uncommon words to generate metadata. The file structure of the LO can also be favorable for AMG algorithms. E.g. the Microsoft Word document format stores footers and headers at specific sections within the file. These sections frequently contain content such as "Page" and page number, author name or file name. The file structure enable the AMG algorithms to avoid these unwanted sections altogether. In this framework the AMG algorithms have the additional advantage of having a reference keyword list populated with course specific entities retrievable from the LMS. Combining this with information from both the LMS and LO regarding the content language, enables AMG algorithms based on Natural Language to operate in a favorable environment for generating LO specific "1.5 Keyword" entities.

7. Creating a SCORM LO

The various algorithms described in the previous chapters have been combined to an executable sequence as presented in "Fig. 2". This research has received inconclusive results regarding if Harvestable metadata from the publishing information and the LO

can result in higher metadata quality for the Extraction algorithm efforts [2, 3]. This is since there has not been sufficient data available in the LOs to evaluate if e.g. the publisher and author of the LO are the same person. The stippled line in "Fig. 2" hence indicates a possibility which has not been verified usable. After the efforts of each AMG algorithm is finished executing, the metadata results are combined to a single metadata record before being included in a Zip-compressed archive consisting of the metadata record and the LO, hence generating a new valid SCORM LO.

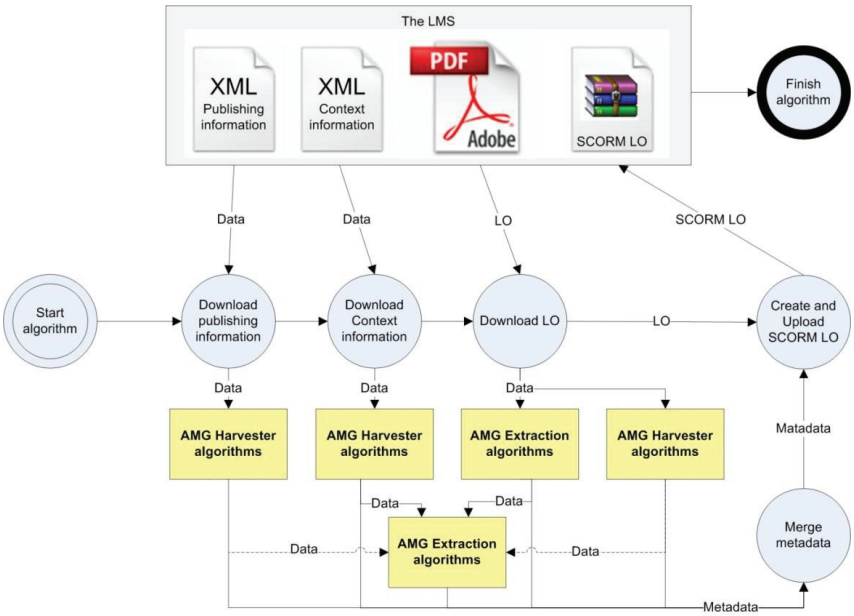


Fig. 1. Framework execution sequence

8. Conclusion

A LMS can be used for much more than to store and publish LOs. It can be used as a distribution channel for contextual metadata. Metadata which in turn can be used by content based AMG algorithms to generate LO specific metadata. This research has presented how existing information sources can be reused in the LMS context as information provider to content based AMG algorithms. By combining usage of AMG algorithms, this research has enabled generation of extensive metadata descriptions to LOs without the need of manual efforts. This while keeping the quality of the generated metadata at a high level.

Inclusion of such technology as part of the LMS would enable easy creation of SCORM

LOs. This framework does not aim at generating all the metadata which is desired present within a SCORM LO. The richness of such descriptions can only be generated in a cooperation with manual efforts by the LO creator or publisher. This framework aims to revert the manual effort to the elements where the manual efforts are needed, taking away elements which do not need the full manual attention.

Similar logics can be used to populate other metadata schemas used in e.g. a company intranet.

Several logistical steps can be included in this framework as part of future work. Firstly, relations between course specific SCORM LOs can be automatically generated on a scheduled basis, including new SCORM LOs as they are published. Secondly, the LO publisher or creator can be made active in the SCORM LO creation process by enabling editing of generated metadata and adding manually created metadata entities. The lecturers and other course responsible personnel can also be actively encourage to generate further course- and organizationally adapted metadata, such as extended keyword lists. Thirdly, a larger scale, practical implementation of the framework would enable more practical feedback of such a solution, possibly also enabling verification of using Harvestable metadata sources for extracting efforts. Fourthly, analysis of actual usage of the SCORM LOs would be in place in order to evaluate if students and lecturers actually desire to take advantage of this data source or if they still prefer the old fashion way of manually creating new LOs from scratch each time there is a need for it.

References

[1] Hämäläinen, M., Whinston, B. A.and Vishik, S., "Electronic markets for learning: education brokerages on the Internet", Communications of the ACM archive, Volume 39, Issue 6 (June 1996), pp. 51 – 58, ISSN:0001-0782, http://portal.acm.org/ft_gateway.cfm?id=228513&type=pdf&coll=portal&dl=ACM&CFID=33796525&CFTOKEN=11974524

[2] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T. and Trætteberg, H., "Using the structural content of documents to automatically generate quality metadata". Webist 2009, March 23-26, 2009. Springer

[3] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T. and Trætteberg, H. "Automatically generating high quality metadata by analyzing the document code of common file types", 2009. JCDL, ACM 978-1-60558-322-8/09/06.

[4] Cardinaels, K., Meire, M. and Duval, E., "Automating metadata generation: the simple indexing interface", *Proceedings of the 14th international conference on World Wide Web*, Chiba, Japan, 2005, pp.548-556, ISBN:1-59593-046-9

[5] Greenberg, J., "Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications". Journal of Internet Cataloging, 2004, 6(4): 59-82.

[6] Meire, M., Ochoa, X. and Duval, E., "SAmgI: Automatic Meta-data Generation v2.0". *In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007*, 2007, pp. 1195-1204, Chesapeake, VA: AACE

[7] Duval, E. and Hodgins, W., "Making metadata go away: Hiding everything but the benefits". Keynote address at DC-2004, Shanghai, China, 2004

[8] It's learning. "It's learning". 2009. http://www.itslearning.com

[9] Edvardsen, L.F.H. and Sølvberg, I.T., "Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment". Webist 2007, March 3-6, 2007. Springer

[10] IEEE LTSC, "IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata", WG12: Related Materials, 2005, http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

[11] DCMI, "Dublin Core Metadata Element Set, Version 1.1." Dublin Core Metadata Initiative, 2008, http://dublincore.org/documents/dces/

[12] Friesen, N., "Final Report on the "International LOM Survey"", CAC JTC1/SC36 Document No: 36C087 2004-08-25, 2004 http://jtc1sc36.org/doc/36N0871.pdf

[13] IEEE LTSC, "LTSC Home Page – IEEE Learning Technology Standards Committee", 2009, http://ieeeltsc.org/

[14] IMS, "Welcome to IMS Global Learning Consortium, Inc.", 2009, http://www.imsglobal.org/

[15] ARIADNE, "ARIADNE Foundation for the European Knowledge Pool", 2009, http://www.ariadne-eu.org/

[16] CETIS, "UK LOM Core v 0.2", 2004, http://www.cetis.ac.uk/profiles/uklomcore/uklomcore_v0p2_may04.doc

[17] eStandard, "Norsk LOM-profil – NORLOM. Versjon 1.0", 2005, http://www.estandard.no/norlom/v1.0/NORLOM_v1_0_mars_2005.pdf

[18] FREPA, "frepa.blog - On e-learning and Learning Technology » Blog Archive » SWE-LOM: a Swedish LOM application profile", 2006, http://www.frepa.org/wp/2006/06/28/swe-lom-a-swdish-lom-application-profile/

[19] CanCore, "Guidelines", 2004, http://www.cancore.ca/en/guidelines.html

[20] Li, Y., Dorai, C. and Farrell, R., "Creating MAGIC: system for generating learning object metadata for instructional content", *Proceedings of the 13th annual ACM international conference on Multimedia*, Hilton, Singapore, 2005, pp.367-370, ISBN:1-59593-044-2, http://portal.acm.org/ft_gateway.cfm?id=1101227&type=pdf&coll=GUIDE&dl=GUIDE&CFID=1299051&CFTOKEN=15183150

[21] ADL, "Sharable Content Object Reference Model (SCORM) 2004 3rd Edition Documentation Suite", 2006, http://www.adlnet.gov/downloads/AuthNotReqd.aspx?FileName=SCORM.2004.3ED.DocSuite.zip&ID=237

[22] ADL, "Advanced Distributed Learning", 2009, http://www.adlnet.gov

[23] Open Archives Initiative. 2004 Protocol for Metadata Harvesting – v.2.0., 2004 http://www.openarchives.org/OAI/-openarchivesprotocol.html

[24] Seymore, K., McCallum, A. and Rosenfeld, R., "Learning hidden Markov model structure for information extraction.", *Proc. of AAAI 99 Workshop on Machine Learning for Information Extraction*, pages 37-42, 1999.
[25] Greenstone. Source only distribution. http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected), 2007
[26] Bird, K. and the Jorum Team. "Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems.", 31st March 2006, Version 1.0, Final, Signed off by JISC and Intrallect July 2006.
[27] Google. "Google". 2009. http://www.google.com
[28] Scirus. "Scirus – for scientific information", 2009. http://www.scirus.com
[29] Yahoo. "Yahoo!", 2009. http://www.yahoo.com
[30] Singh, A., Boley, H. and Bhavsar, V.C, "LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology." National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004
[31] Giuffrida, G., Shek, E. C. and Yang, J., "Knowledge-Based Metadata Extraction from PostScript Files.", Digital Libraries, San Antonio, Tx, 2000, ACM 1-581 13-231-X/00/0006
[32] Kawtrakul A. and Yingsaeree C., "A Unified Framework for Automatic Metadata Extraction from Electronic Document.", *Proceedings of IADLC2005*, 25-26 August 2005, pp. 71-77.
[33] Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M., "Automated Template-Based Metadata Extraction Architecture", ICADL, 2007.
[34] Li, H., Cao, Y., Xu, J., Hu, Y., Li, S. and Meyerzon, D. "A new approach to intranet search based on information extraction", *Proceedings of the 14th ACM international conference on Information and knowledge management*, Bremen, Germany, Pages: 460 – 468, ISBN:1-59593-140-6, 2005, ACM New York, NY, USA.
[35] Liu, Y., Bai, K., Mitra, P, and Giles, C.L., "TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries." JCDL'07, June 18–23, 2007, Vancouver, Canada, ACM 978-1-59593-644-8/07/0006
[36] Boguraev, B. and Neff, M., "Lexical Cohesion, Discourse Segmentation and Document Summarization", 2000, RIAO.
[37] LOMGen. LOMGen, 2006. http://www.cs.unb.ca/agentmatcher/LOMGen.html
[38] Greenberg J., Spurgin, K., Crystal, A., Cronquist, M. and Wilson, A., "Final Report for the AMeGA (Automatic Metadata Generation Applications) Project", 2005, UNC School of information and library science.
[39] Li, Y., Dorai, C. and Farrell, R., "Creating MAGIC: system for generating learning object metadata for instructional content", *Proceedings of the 13th annual ACM international conference on Multimedia*, Hilton, Singapore, pp.367-370, 2005, ISBN:1-59593-044-2
[40] Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekema, A., McCracken, N.J., Silverstein, J. and Sutton, S.A., "Automatic metadata generation and evaluation", *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval, August 11–15, Tampere, Finland, 2002. ACM Press, New York, pp.401–402.

[41] Jenkins, C. and Inman, D., "Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF", 0-7695-1022-1/01, 2001 IEEE

P6: Creating Metadata is a Costly Manual Process – And it can be Automated

Published in: "Digital Libraries and Knowledge Organizations."
Macmillan Publishers India Ltd., pp. 356-362, 2012. ISBN 978-935-059-076-8

Ingeborg Torvik Sølvberg[1] and Lars Fredrik Høimyr Edvardsen[2]

[1] Norwegian University of Science and Technology
NO-7491 Trondheim, Norway
Ingeborg.Solvberg@idi.ntnu.no

[2] Norwegian University of Science and Technology/
NO-7491 Trondheim, Norway
and
EDB ErgoGroup
NO-0483 Oslo, Norway
Lars.Fredrik.Edvardsen@edb.com

Abstract: Enabling efficient retrieval and re-usage of digital documents is a major challenge as many documents on the Internet and on Intranets are poorly described with metadata. Manual generation of quality metadata requires skilled human resources, it is costly and time-consuming. As a result, metadata related to the documents are too often insufficient or even in-correct. Automatic Metadata Generation (AMG) algorithms could perform metadata generation efforts in seconds without the need for human efforts. This can open for increased performance of e.g. search engines using document metadata as a data query source and in the query results. This paper presents characteristics of AMG algorithms, three different document collection environments and how non-visual modification of an organization's document templates can improve the efforts of AMG algorithms.

Keywords: Automatic metadata generation, AMG, Content Analysis and Indexing

Is not included due to copyright

SP1: Use of It's learning at NTNU – A Quantitative and Qualitative study

Original title in Norwegian: Bruk av It's learning ved NTNU – en kvantitative og kvalitativ studie

Internal study report at NTNU, conducted by the Program for Learning with Information and Communication Technology (Program for Læring med IKT (LIKT)) in order to review usage of It's learning at NTNU. pp. 1-157. January 2008. Published at and by NTNU in January 2008.

The report is written in English. The abstract below has been translated into Norwegian for use in this thesis.

Line Kolås [1], Lars Fredrik Høimyr Edvardsen [2] and Leif Martin Hokstad [1]

[1] Dept. of Education
Norwegian University of Science &Technology
NO-7491 Trondheim, Norway
line.kolas ; leif hokstad @svt.ntnu.no

[2] Dept. of Computer and Information System [2]
Norwegian University of Science &Technology,
NO-7491, Trondheim, Norway
lars.edvardsen@idi.ntnu.no

Summary: The goal of this study were to present a timeframe that shows how It's learning is used at NTNU in the spring of 2007. Two quantitative studies and one qualitative study have been conducted. The available data set is mainly related to teacher initiated activities at It's learning. It shows to a large extent how It's learning was used in the spring semester of 2007 in numbers and tables from statistics generated in It's learning, but also with reflections and considerations by the interview subjects at the various faculties.

The Quantitative study (part 1) presents use of It's learning at NTNU in the spring of 2007 in numbers, diagrams and tables. The data set is collected from statistics generated in It's learning at NTNU. The results are presented as total use of It's learning at NTNU, use of It's learning in the various course types (introduction subjects, elementary subjects, advanced subjects, master subjects, PhD subjects and

EVU[17] subjects) and use of It's learning at the various faculties and institutes at NTNU.

The Quantitative study (part 2) show what types of file formats and document types that is hiding behind the phrase "files" in It's learning.

The Qualitative study is based on 18 face-to-face interviews with teachers at various institutes at NTNU, supplier presentations and one student interview. The interviews had as main goal to get the interview objects to reflect upon the pedagogical usage of It's learning in their own teaching in the spring of 2007.

The study show that It's learning is being used more as an administrative tool than an educational system at NTNU. The interviews with the teachers showed their view if It's learning as a course administrative system rather than a system for promoting learning. This is contrary to the supplier's view of the system and their focus of it being a pedagogic system.

The functionality of It's learning is limitedly used. Basically It's learning is used as an administrative tool to ensure information flow to the students and to publish static files. The quantitative study show that just above 50% of the courses has uploaded at minimum 1 file to It's learning, while 39% of the courses had uploaded more than 10 files. 30% of the courses have uploaded on average 1 file per week and only 16% of the courses had uploaded on average 2 or more files per week during the spring semester of 2007. There is also a general tendency that if no files are uploaded, then none of the other tools in It's learning are being used. It is therefore possible to suggest that about 30% of the subjects have a regular use of It's learning, even if it's difficult to define "regular use".

The tools of It's learning is limitedly used. Forum, notes, links and task tools are used by about 20% of the courses at NTNU, while surveys, tests, text collections, explanatory sequences and conference tools are being used by between 7% and 0.2% of subjects. The figures show little use of the discussion forums and conferences, which suggest that It's learning is not used as a two-way communication solution by NTNU. The quantitative study also shows that there is not a correlation between extensive use of specific It's learning tools and usage of the other tools available in It's learning.

The quantitative study also shows that all faculties have introduced It's learning to a greater or lesser degree, and that only a small number of courses have not been using It's learning in the spring of 2007. Over 60% of the courses at the NT[18] and SVT[19]

---

[17] EVU = Etter- og videreutdanning. After and extended education

[18] NT = Fakultet for naturvitenskap og teknologi.
The Faculty of Natural Sciences and Technology.

[19] SVT = Fakultet for samfunnsvitenskap og teknologiledelse.
Faculty of Social Sciences and Technology Management.

faculties were activity using It's learning. The lowest frequency of use were at the DMF[20] and AB[21] faculties with less than 40% of their courses using It's learning.

The courses at NTNU vary extensively in terms of educational goals, the number of students and whether students are on campus daily, or whether they are distributed across the country. In addition, some courses have students out in practice for much of the study period. In regards to It's learning tool usage (notes, links, tasks, text collections, explanatory sequences and conferences) are most commonly used at EVU subjects, while the discussion groups are most commonly used in the basic courses (bachelor) and tests are on average mostly used in PhD courses.

The qualitative study focused on the pedagogical principles of variation, individualization, differentiation and meta-learning. Variation was considered as an important educational principle, but first and foremost in the auditorium, not on It's learning. With respect to usage of pedagogical methods used, usage of presentations is the most common with contents of text, images and video based presentations. In addition, exploration / problem solving is used to some extent. The other methods, such as games, simulations and cooperation solutions are seldom used. The study also shows that It's learning is not used as a PLE[22] to individualize and differentiate instruction. This can probably be justified in that the system is not designed as a PLE, but also because the staff at NTNU in part does not want to individualize instructions at the university level and that they do not see opportunities to use It's learning to contribute to such learning activities.

No departments at NTNU have guidelines for a common menu structure (tree navigation structure). The interviews reviled that many courses have menu structures that are not very well planned and thought out. Menu structures were partially chronological, media-based and thematically structured. Some described that both teachers and students had difficulty finding information and data in It's learning.

Interviewees had difficulty describing missing functionality in It's learning, although specific features such as synchronous conferencing tools (with enable application sharing and video conferencing) and formula editor was called for. Most found it easier to criticize certain features of the current system. The interviews showed that It's learning was seen as a somewhat unstable system. It was also clear that some lacked confidence in the security and uptime of It's learning.

Lack of resources / lack of time can be regarded as partial reason for why It's learning is not used more in many courses, but the view of It's learning as a subject-administrative system also limits the teachers looking for educational opportunities within the system.

---

[20] DMF = Det medisinske fakultet ved NTNU. Faculty of Medicine.

[21] AB = Fakultet for arkitektur og billedkunst. The Faculty of Architecture and Fine Art

[22] PLE = Personal Learning Environment

SP2: Using Document Code to automatically generate high quality metadata: An Auditing case study

Lars Fredrik Høimyr Edvardsen
Intelligent Communication AS /
Norwegian University of Science and Technology
Akersgata 49
NO-0180 Oslo, Norway
+47 986 90 441
edvardsen@hotmail.com

Ingeborg Torvik Sølvberg
Norwegian University of Science and Technology
Sem Sælands vei 7-9
NO-7491 Trondheim, Norway
+47 73 59 60 27
ingeborg.solvberg@idi.ntnu.no

ABSTRACT: Enabling efficient retrieval and re-usage of digital documents is a major challenge as many documents on the Internet and on Intranets are poorly described with metadata. Manual generation of quality metadata requires skilled human resources, is costly and time-consuming. As a result, metadata related to the documents are too often insufficient or even in-correct. Automatic Metadata Generation (AMG) algorithms could perform similar metadata generation efforts in seconds without the need for human efforts. Recent research indicates that the document code of Word documents can be a basis for automatically generating high quality document metadata without the need for human interaction. This paper puts this to the test as high quality metadata is attempted retrieved from a range of actual auditing documents. This paper also shows how optimizing the document templates can vastly improve the quality of the generated metadata using a single AMG algorithm even when the document collection contains extensive diversities.

Categories and Subject Descriptors
H 3.1 [Information Systems] Content Analysis and Indexing – abstracting methods, indexing methods

H 3.7 [Information Systems] Digital Libraries – collection

C 3 [Computer Systems Organization] Special-purpose and application-based systems – Real-time and embedded systems

General Terms: Algorithms, Reliability, Standardization, Verification.

Keywords: Automatic Metadata Generation, Metadata, Harvesting, Extraction, Document Code, Metadata Quality, Microsoft Word, OpenXML.

1. INTRODUCTION

Large amounts of scarce human resources are still being used to create similar resources, such as documents [18]; partly because people are not aware of others work, the lack of sharing opportunities, and the inability to retrieve available documents. This situation is, however, dramatically improved the last years due to the large document collections being made available on the internet by publishers,

organizations and individuals. A major challenge for content management in such collections is the generation of high quality metadata. Manual generation of quality metadata requires skilled human resources. Trained librarians and archivists can assist authors to create and publish metadata, but this is a costly and time-consuming process. Automatic Metadata Generation (AMG) is methods for generating metadata without manual interaction using computer program(s) to interpret the document and possibly the document context. AMG is based on the observation that information that equals the desired metadata, directly or indirectly, may already be contained in the documents or in the context as:

- **Visual descriptions:** By viewing the document through its native application or as a print-out, visual characteristics can be seen, such as the paper format and promotion of specific sections (e.g. some text with larger letters).
- **Technical descriptions:** By analyzing technical information from the document or the system in which the document is stored, other characteristics can be obtained, e.g. file size, file format and dates.
- **Intellectual content descriptions:** By analyzing the user specified textual content of a document, the intellectual content created by the user can be determined, such as the actual letters used to stipulate the document title.
- **Context descriptions:** There is commonly a link between the document which is created and the place in which it is published. E.g. published site and publisher role at that site.

The author of the document has hence directly or indirectly specified the desired content of metadata elements. This can be utilized as AMG strive to avoid excessive manual efforts when similar metadata can be generated automatically based on existing data sources [5, 6, 9, 10, 11, 15, 26]. Previous work have documented that analysis of the document file format, the document code, as seen in Figure 1, can reveal extensive amounts of information regarding the document without having to use an interpreter application to re-generate the documents' visual appearance [9, 10, 11]. Hence, specific document content can be uniquely identified and retrieved regardless of the visual appearance of the document and where in the document that the desired content is located. In these researches the studied dataset were based on documents published on the Learning Management System (LMS) at the Norwegian University of Science and Technology (NTNU). The documents were extremely diverse regarding intellectual content and visible presentation, as they were collected from a number of publishers in a vast field of different subjects [9]. Though, even with this diversity, the educational "theme" influenced the documents; they were mainly presentations, exercises and academic papers.

This paper looks to validate conclusions regarding usage of the Document Code for AMG efforts by using a significantly different document collection: Documents created by Auditors for Auditing purposes. These are strictly formatted documents to ensure compliance with the company profile, juridical and professional validity and correctness towards the customer, authorities and other parties of interests, such as unions.

```
<w:body>
  <w:p>
    <w:pPr>
      <w:pStyle w:val="Title" />
    </w:pPr>
    <w:r>
      <w:rPr>
        <w:lang w:val="en-US" />
      </w:rPr>
      <w:t>Automatic Metadata
           Generation</w:t>
    </w:r>
  </w:p>
</w:body>
```

Figure 1: Example of simplified document code from an Open XML (MS Office 2007) document

That said, the documents' visual appearance vary extensively caused by personal preferences of the Auditors, the customer and as a result of Audit issues discovered. As a result, content sections are moved around and the length of each content section varies extensively. In addition, this Auditing firm uses document templates which contain a minimum of visual content section promotion. All these properties are challenges for AMG efforts based on visual characteristics. Though, when basing the AMG efforts on the Document Code these issues should not have an influence on the end resulting metadata.

Chapter 2 presents the State-of-the-Art of the field of AMG while Chapter 3 presents the research setup. In chapter 4 this paper compares usage of AMG effort on Auditing documents against the "Reference Work" from the NTNU LMS as sited in [9]. Document templates can be modified to include content specific Meta tags to improve the efficiency of the AMG efforts based on the Document Code. Chapter 5 analyzes documents created by Auditors based on document templates optimized for AMG efforts. Chapter 6 concludes and presents future work.

2. AUTOMATIC METADATA GENERATION

AMG algorithms are sets of rules for processing of data source(s), identification of desired content, and collection and storage of data in accordance with a metadata schema. AMG can be used on all digital resources ("documents") and works by retrieving consistently present content from the resource. AMG algorithms can use the document itself and the context surrounding the document as data sources. Collecting embedded metadata is known as metadata harvesting [7, 15]. The process by which AMG algorithms create metadata that previously has not existed is known as metadata extraction [1, 19]. AMG efforts represent a balancing act between obtaining high quality metadata descriptions and avoiding the generation of metadata that is incorrect

for the description of the document. Four different approaches are used to generate metadata, as illustrated in Figure 2.

Harvesting of embedded metadata: This approach uses the embedded metadata created by applications or by the user and stored as part of the document [1, 2, 14, 17, 27, 29]. Embedded metadata can also be harvested from contextual information sources, such as Intranets, and used directly or in-directly for AMG purposes [9, 10, 11]. Harvesting of embedded metadata is vulnerable to generating false metadata if the embedded metadata is in-correct.

Extraction based on visual characteristics: This approach uses a content presentation application to create a visual representation of the document before executing rules to extract content based on the visual characteristics of the document [13, 21, 28, 30, 31]. This approach can enable generation of entities to populate a number of elements, though is vulnerable to generating false metadata if the documents do not share the visible characteristics with which the algorithm has been developed to perform. Hence, such algorithms only perform as desired on pre-known document types and not on document is general.

Extraction of metadata based on natural language: This approach uses a content presentation application to retrieve only the intellectual content of the document, creating a plain text data source upon which rules based on natural language are executed [3, 12, 16, 22, 24, 25]. Such algorithms commonly include collection of unique words and comparisons of the document vocabulary against reference ontology for generating keywords, descriptions and subject classification. This approach is vulnerable to generating false metadata if the data sources contain documents in multiple languages, document sections in different languages or contains header or footer fields since the text from these fields are presented on every page hence occur frequently.

Extraction based on document code analysis: This is a new approach presented by this project. This approach uses analysis of the code of the document directly without the need for additional content presentation applications to interpret the document content [9, 10, 11]. This enables full and direct access to the entire document's content. This includes template identification, template content identification and formatting characteristics regardless of visual characteristics, and the language of the intellectual content. It also enables effective usage of other AMG approaches when the document code does not provide the desired data. This approach requires that the document code is understandable for the AMG algorithm and that the end-user does not misuse the document template. In an analysis of Conference and Journal papers, this showed to be an issue, as document creators did not follow the specified paper templates [8].

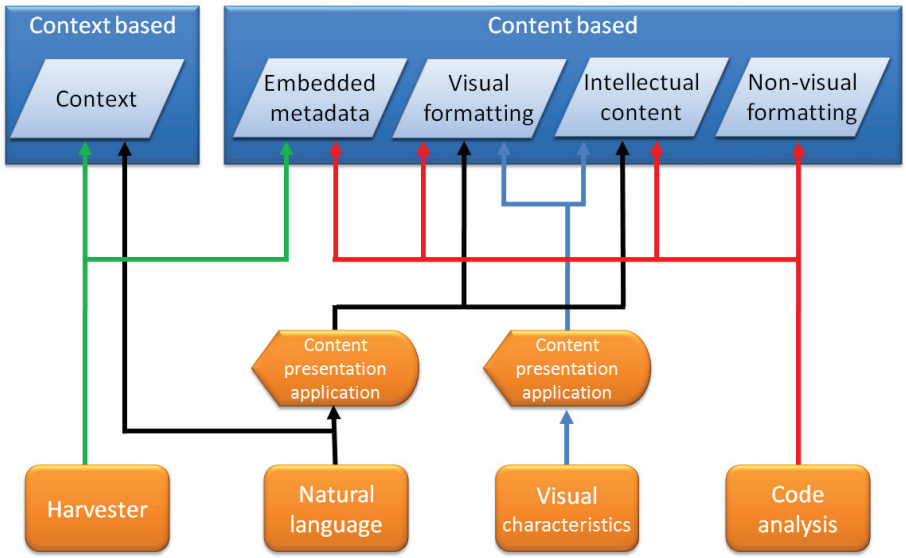


Figure 2: AMG analysis algorithms and the data sources which they use.

Finding objective criteria to validate the results of AMG algorithms is a challenge in it selves. The research results were evaluated using a framework for measuring "quality" presented by [23]. In the context of this paper, this framework categorizes "quality" based on:

- **Syntax:** Analysis of the document formatting to see that it complies with the document's format standard.
- **Semantics:** Analysis of the entities presented to see if they are valid and in accordance with the document format's relevant metadata schema.
- **Pragmatics:** Analysis to determine if the user-interpreted properties are reflected in the metadata.

Additionally, supplemental quality terms were used based on [4]. This framework supplement [23] by including dedicated metadata quality terms for:

- **Completeness:** Completeness reflects two issues: (1) The use of as many elements as possible; and (2) that the user's desired elements are present in the metadata records.
- **Accuracy:** The entities should describe the document correctly and factually.
- **Provenance:** There should be a record of who created the metadata.
- **Conformance to expectations:** Assumes that the users' expected elements are available.
- **Logical consistency and coherence:** Logical consistency relates to compliance with the local metadata schema. Coherence relates to whether the elements are made available.
- **Timeliness:** Timeliness relates to two issues: (1) Currency: when the document changes while the metadata remain unchanged. (2) Lag: when the document is disseminated (distributed) before some or all metadata is knowable or available.

- **Accessibility:** The metadata are available and understandable to the users.

The quality scale is measured subjectively as: (1) Very high: The dataset can confirm a high degree of correctness, (2) High: The dataset can confirm a high degree of correctness, although more than a few exceptions were discovered, (3) Undeterminable: The dataset could not verify either correct or false entities for the given element, so that a conclusion could not be drawn, (4) Low: Systematic false entities were verified to be present, (5) Very low: An extensive number of false entities were verified as present in the dataset.

3 RESEARCH SETUP

This paper will follow in the footprints of the Reference Work by using an identical research setup, though with another document collection: This paper gained access to 150 official Auditing documents in the MS Word document format retrieved from an Auditing firm. Non-official auditing documents, correspondence and draft documents were not made available to this paper. The documents were lossless converted into the MS Word 2007 Open XML document format before analysis. Metadata harvesting and extraction efforts were conducted using the same State-of-the-art AMG algorithms as the Reference Work were based upon. The analysis results are compared using the same framework for referring to "quality".

4. AUTOMATICALLY GENERATING METADATA FROM AUDITING DOCUMENTS

This chapters starts by presenting what is "right" and "wrong of Auditing metadata in Chapter 4.1. The embedded metadata is then analyzed in Chapter 4.2 before the results of the metadata extraction efforts are presented in Chapter 4.3.

4.1 "Right" and "Wrong" of Auditing metadata

These are strictly regulations regarding the content of auditing documents in order to ensure compliance with the company profile, juridical and professional validity and correctness towards the customer, authorities and other parties of interests, such as unions. As a result, a number of people can be involved in the creation of a single Auditing document. In addition, secretaries are used to write documents based on the Auditors' notes. Still, there is only one specific Auditor who should be registered as the document author. Similarly, there is commonly a specific date which specifies the document creation date, regardless of the actual date of creation. Due to such issues, the studied organization bends the rules of right and wrong metadata entities as there is a distinction between technically and juridical specified "right". Of these, it is only the juridical correct metadata which is of value to the organization.

4.2 Embedded metadata

Common document creation applications automatically generate extensive metadata descriptions. These include an author field, created date and a document title. Such metadata generation is commonly performed as long as the user does not explicitly specify that metadata should not be generated. Documents therefore commonly contain extensive metadata descriptions, often created without the user being aware of this. There can be a number of different sources for generating such metadata, including the document it selves, template information, system information or information gathered from external data sources.

There are issues regarding the embedded metadata from the documents in the dataset. None of the documents contained meta-metadata which specify who have created the metadata: The original application, a third party application, the template creator or end user. The provenance aspect of the metadata quality was hence very low. These observations are identical to the Reference Work.

Gaining access to embedded document metadata is commonly not straight-forward as a result of file formats using their own approach to embedded metadata. The metadata harvesting efforts therefore needed to be adapted to each document format in order to access, interpret and retrieve the metadata. This reduces the quality of the accessibility of the metadata. It also requires ongoing efforts to adapt the harvesting efforts to new document formats or new versions of the document formats over time. No documents were discovered with syntactic false formatting. In terms of syntax, this paper has confirmed that the embedded metadata is of high quality, as no document contained falsely syntactical content. The entities were of high semantic quality as all elements contained valid entities. In terms of pragmatic there was however a number of issues caused by a number of issues, resulting in entities which users would interpret as false. These issues will be presented in the following chapters.

4.2.1 The Title element

The Title element is automatically populated with an entity the first time a Word document is stored. Though, if a title already exists, e.g. from the document template, then a new title is not added or updated. This can cause timeliness issues. There are also issues regarding what content that the application selects as title for the document.

In the Reference Work only 14% of the embedded Title entities were identical to the content which was manually recognized to be the documents' title, referred to as the documents' visible title. In this dataset 56% of the embedded Title entities were identical to the documents' visible title. The false entities were commonly caused by false content selected by the MS Word application to generate the Title entity. Up until recently, the Title entity was populated by content on the first line within the document. As presented later in chapter 4.3.1, this is not an optimal approach. Current versions of the MS Word application do not generate Title entities.

In terms of pragmatics, the embedded Title entity is of a low quality if present. If not present, the pragmatic quality is very low due to completeness, accuracy and conformance to expectations. This confirms the results of the Reference Work.

4.2.2 The Creator element

The Creator element is also commonly automatically populated with data retrieved from the user's user name or retrieved from the document template. The Reference Work reported issues validating the Creator element, as only 22% of the documents contained visible author information.

All official Auditing documents must by law present the document author. This organization has a strongly enforced policy regarding usage of descriptive user names. As a result, the vast majority of documents contained either a full name or abbreviations indicating a specific person within the organization. In 10% of the embedded metadata entity could be interpreted as the visible author, commonly specified by a full name, last name or affiliation. In the remaining 90% of the documents the embedded metadata referred to another person than the documents' visible author.

The embedded metadata refer more to the technically correct creator, than the pragmatically user expected juridical creator. However, as there was commonly timeliness issues due to template entities passed on to the resulting document, the quality of technically correct author was low. In terms of pragmatics, the embedded Creator entity was of very low quality caused by the template issue described above, plus that a large number of documents were written by a secretary rather than the juridical author.

4.2.3 The Date element

The Date element is commonly automatically populated with data from the clock of the local computer in which the document were created on. As local clocks can go wrong, the resulting entities cannot be fully trusted if not verified by another data source. The Reference Work experienced extensive issues regarding lack of content which the Date entities could be validated against.

All official Auditing documents must by law present the current date. As such, the embedded entities could be compared against a visible date present on all documents. Though, only a handful of the documents had embedded entities and a visible date which specified the same date. These results are strongly influenced by Auditor documents that *should* be created on a specific date. Hence, the visible date refers to the juridical date while the embedded entities refer to technical dates.

As with the Creator element, the embedded metadata hence refer to other data than what the users are expecting. In terms of pragmatics, the embedded Date entities are

of very low quality. Technically speaking, no entities were discovered that indicate vastly false entities, e.g. wrong year or month. Technically, the embedded Date entities are of high quality.

4.2.4 The Template element

The Template element is not a common part of metadata schemas. Though, in order to enable efficient AMG algorithms based on visible characteristics, determining the based template is of high value [9]. This since each document type can contain different amounts of content which is desired retrieved and visual appearance and location of this content.

In the Reference Work ninety-five percent of Word documents were based on the blank default template. This template is commonly without any content and hence do not contribute with descriptive information regarding the document.

In this dataset eighty percent of the documents were based on a template other than the blank default template. The name of these templates indicates usage as a specific Auditing document type. In only thirty percent of the documents the indicated usage area and the actual document content were identical. Hence, the same document template is used for a number of different document types.

Nineteen of twenty documents referring to the blank default template had document content with visible appearance and title identical to the template specified documents. This indicate that more than one template is used to generate the same type of documents.

In terms of pragmatics, the embedded Template entity was of very high quality as no false entities were discovered. Though, in terms of Accuracy, the quality were very low, as only a very limited number of document types could be identified based on the present entity.

4.2.5 The Language element

AMG algorithms based on natural language processing are adapted to handle one specific language of the intellectual content. Hence, determining the language of the intellectual content is of essence for such algorithms when they are executed in a multi-linguistic document environment.

In the Reference Work no documents contained a harvestable Language element. This result is identical to the observations this paper have experienced. This confirms the Related Works conclusion that the metadata quality in terms of completeness was hence very low.

4.3 Extracted metadata

In the Reference Work the Title, Creator and Language elements were attempted automatically generated using extraction efforts. The Date and Template elements were not attempted generated as the dataset did not contain data to build algorithms upon. This paper will re-test the used algorithms on the Auditing dataset.

4.3.1 The Title element

Specifying the correct document title is essential for efficient document identification and retrieval for the Auditors. The embedded Title was not of a quality adequate for the Auditors, hence showing a need for other AMG efforts.

All but one document templates in this study contained a visible title field. Though, their visual appearance varied extensively, as some titles were promoted with large letters on the top of the page, to titles further down on the page in plain text as on formal letters. The Reference Work uses a number of AMG algorithms to automatically generate Title elements:

A. **File name:** Harvested from the file system [2].
B. **Embedded metadata:** Harvested from the document [14, 17, 20, 28 30, 31].
C. **First line:** Extracted from the first visible line of text [17].
D. **Largest font:** Extracted the text section on the first page based on the largest font size [13, 14].
E. **Document Code:** Extracts title section specified in the document code [9].
F. **Document Code plus usage of the largest font, context filter and alternative data sources:** Extracts title section specified in the document code. If not present, then the content with the largest font is attempted retrieved, followed by the embedded tile and file name. Known false data were excluded as part of the execution efforts [9].

Table 1 presents how the AMG algorithms presented on the LMS and on the Auditing datasets. In the Reference Work each algorithm performed correctly or partly correctly for more than half of the documents, with the exception of plain Document Code based efforts. Usage of the document content section with the largest font on the first page of the document showed to generate the highest number of correct titles with its correctness rate of sixty-nine percent. None of the Word documents in the LMS dataset contained Meta or Style tagged sections that indicated a title. Hence, usage of the document code exclusively resulted in no titles found and no entities created. The Reference Work showed how the Document Code approach could be combined with the other AMG approaches and content filters. This content filter excluded content such as the course name (compared against the published location on the LMS). When combining these AMG approaches, as much as ninety-one percent of the titles could be completely and correctly extracted.

Table 1: Usage of AMG algorithms to generating Title entities

Title Algorithm	Correct		Partly correct		No result		False	
	Ref.Work	***Auditing***	***Ref.Work***	***Auditing***	***Ref.Work***	***Auditing***	***Ref.Work***	***Auditing***
A. File name	40%	0%	45%	80%	0%	0%	15%	20%
B. Embedded	27%	0%	29%	0%	8%	0%	36%	100%
C. First line	38%	0%	15%	0%	1%	0%	46%	100%
D. Largest font	69%	20%	8%	0%	1%	0%	22%	80%
E. Document Code	0%	22%	0%	0%	100%	78%	0%	0%
F. Document Code extended	91%	28%	6%	5%	0%	0%	3%	0%

In the Auditing dataset these algorithms performed differently. Algorithms A to D performed poorly due to a mismatch between the algorithms desired content and the documents embedded and visible appearance: Algorithms A and B commonly retrieved the document type or template name, while algorithms C and D retrieved the company or Auditor name.

Twenty-two percent of the Auditing documents contained a Meta or Style tagged Title section which were identified and extracted using algorithms E and F. All of these titles were complete and correct titles. Most of these documents were based on the same templates or document type. Feedback from the Auditors indicates no intentional usage of Meta tags. Rather, such tags were included in these documents because selected templates contained a Meta tagged title and the Auditors frequently corrected this title rather than creating a new title. Why such Meta tags had been included in some templates were not known by the Auditors.

When moving from algorithm E to F, the Reference Work experienced an increased correctness rate from zero percent to ninety-one percent. In the Auditing dataset the same algorithms generated an increase in correctness rate from twenty-two to twenty-eight percent. This time the advanced Document Code based algorithm (algorithm F) was not able to gain benefits from correct or partly correct entities generated by the supplementary AMG algorithms. This was due to the content filters executed between sub-AMG algorithms. These filters were set up to exclude course codes and the publisher name from the generated entities retrieved from the LMS. Similar information was not available for the Auditing documents. Hence, the AMG algorithm did not have content to exclude. When completing the first two stages of

the advanced Document Code algorithm, all documents were given a valid title either based directly on the Document Code or on the Largest Font sub-algorithm. The algorithm were not able to take advantage of the First line, Embedded metadata or File name data sources. Though, since neither of the algorithms A to D provided substantial positive results, enabling usage of these data sources would not significantly increase the performance of algorithm F.

Due to the low quality in terms of Accuracy, the pragmatic quality of all these algorithms was low.

4.3.2 The Creator element

As the juridical correct author was not specified using metadata harvesting, other approaches would be needed in order to gain de desired metadata quality.

Using extraction methods for populating the Creator element proved to be an extra challenge within the LMS case study [9]. This since only a small selection of the documents contained visible author information. As a substitute, a Publisher element was generated based on the LMS' user login information. Such publishing information is not available in this case study since the documents are shared though a common file storage.

No documents used the author name as part of the file name making algorithms like algorithm A for generating Title entities without value. To enable efficient algorithms based on algorithm C and D for generating Title entities, visual consistency within the documents is essential.

Visually, the desired Auditor name can be present in a number of different locations: On the first line, on the second line or one of the last sentences on the last page of the document. This diversity is a challenge for AMG algorithms based on visible characteristics. Without positive identification of the document type, 25% of the Author names could be retrieved from the document collection, as one in four templates contained author name on respectfully the last, second to last or sixth from last line on the last page. The remaining 25% of the templates had author names printed inside of the document; normally close to the middle of the page, though not stationary in one location.

A 25% chance of correctness is actually higher than the correctness rates gained by the Reference Work. Here, basic approaches for extracting the first, second and last line resulted in correctness rates between 0% and 3%.

Still, the generated entities would in general be of a very low pragmatic quality as the accuracy and conformance to expectations would be very low.

Author information can be Meta tagged for identification and retrieval by using the Document Code. This way extraction effort could be performed regardless of the documents' visible formatting. None of the documents contained such Meta tags that

identified a Creator content section. Extraction efforts based on the Document Code would hence not result in any entities, leaving the pragmatic quality at a very low level. These results are identical to the Reference Work.

4.3.3 The Language element

None of the Auditing documents contained embedded Language elements. The Reference Work states that the language of the intellectual content can be identified suing Meta tags stored as part of the document in order to enable application services such as spelling and grammar control.

All the Auditing documents contained US English language Meta tags, though one of these documents *used* these tags. This confirms the Reference Work that extraction efforts need to be focused on the tags that are in practical use and not solely on their presence.

Language Meta tags were used to describe content in Norwegian and New-Norwegian. This paper can confirm that all the text sections were formatted with a single language tag and that these Meta tags specified the correct language.

In terms of pragmatic quality, the generated entities were of a very high quality, sectioning the documents into sub-sections with an accurate language specification with full conformance to expectations. This confirms the Reference Work that a Document Code based efforts would be highly valuable for AMG algorithms based on natural language processing as each of the documents' subsections are identified to contain intellectual content of a specific language.

5. EXTRACTING METADATA FROM DOCUMENTS BASED ON MODIFIED TEMPLATES

Neither the harvesting or extraction efforts resulted in desired entities for the Title and Creator elements: The harvestable entities were based on technical data and not juridical data while the extraction efforts were not able to identify the desired sections. A survey among the Auditors revealed that very few of the Auditors were aware of the embedded document metadata. Manual creation or editing of metadata was viewed as a waste of productive time. Hence, none of the Auditors were willing to spend their own time on creating metadata. Embedded metadata would hence not be populated and updated even though the Auditors are aware of their presence.

5.1 Creating new templates

All the Auditors based new Auditing documents on existing documents or templates. Hence, there is always a structured document which is the basis for new documents.

This paper retrieved a dozen Auditing document templates which were in use and included Meta tags to identify content of special interest to the Auditors. Meta tags were created to identify the document tile, juridical author, customer name and juridical date. The visible appearance of each template was not changed in any way.

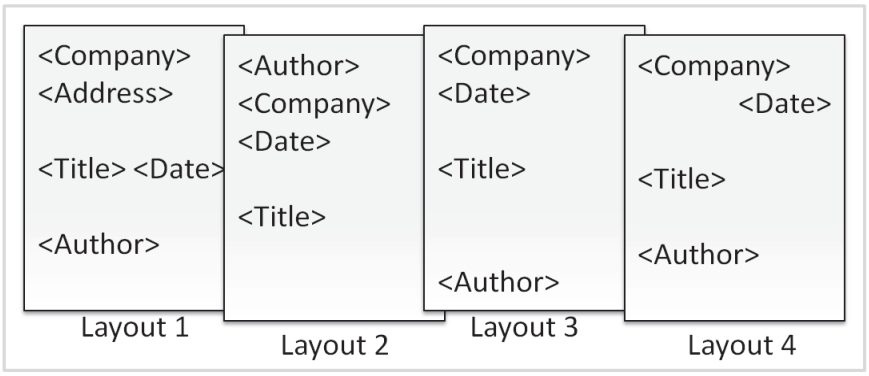


Figure 3: Example of different visual layouts of content sections in the document templates

The included Meta tags were identical for all the templates regardless of the documents' visible layout. Figure 3 illustrates where different content sections are located on four of the templates. In addition, various font types, sizes and visual promotion distinguish the documents from each other. All the templates do not include the same sections. The single structured Meta tags among all the templates enable a single AMG algorithm to identify the desired document section using Meta tags regardless of where in the document the content is present and of the visible appearance of the document. Hence, the same AMG algorithm could be used for extracting the desired content from all the different document templates.

The templates were re-introduced to the Auditors to use during their next auditing efforts. The Auditors were not given any training or instructions describing how the updated templates "should" be used. Their usage is hence based on own preferences.

5.2 New extraction results

After having the updated templates in use for a few months, this paper received a few dozen documents based on the updated document templates. Extraction efforts based on the Document Code were enforced in order to retrieve desired content as high quality metadata.

All the titles and customer names were completely and correctly identified and retrieved from the new documents. The various visual appearances of the content sections that contained the desired content and their location within the document did not have any effect on the metadata extraction results.

The Date and Author sections were not completely correct identified and retrieved. "Only" 94% of the Author names and 89% of the Dates were completely and correctly retrieved. Further analysis of the documents revealed the cause: In all the documents without perfect results, the Date section was located on the line above the Author section and both sections were using the same visual formatting. Behind the visual appearance, these documents did not contain Meta tags identifying the Date section, even though this was specified in the used template. Hence, the section which was Meta tagged had been deleted. These documents had lower metadata quality due to lack of accessibility to the desired content section. The Author sections were present in all the documents. However, occasionally the Author Meta tags covered content that should have been Meta tagged as Date. As a result the extracted Author entity did not always conform to expectations, lowering the metadata quality.

5.3 Comparing documents based on the original and the updated templates

There is a clear distinction between documents created by the original templates and the updated templates in terms of the documents' Document Code. This is even though the documents' visible appearance is virtually identical.

The AMG algorithms which generated the most correct results described in Chapter 4.3 managed to generate correct entities for a quarter of the documents. These were algorithms based on various visible characteristics. Using the same algorithms on the updated dataset would generate identical results. Chapter 5.2 has shown that AMG efforts based on the Document Code would result in entities of vastly higher quality. This since the Meta tags identifies the document sections that are of interest for that specific metadata element, sections where the author(s) have specified the desired content.

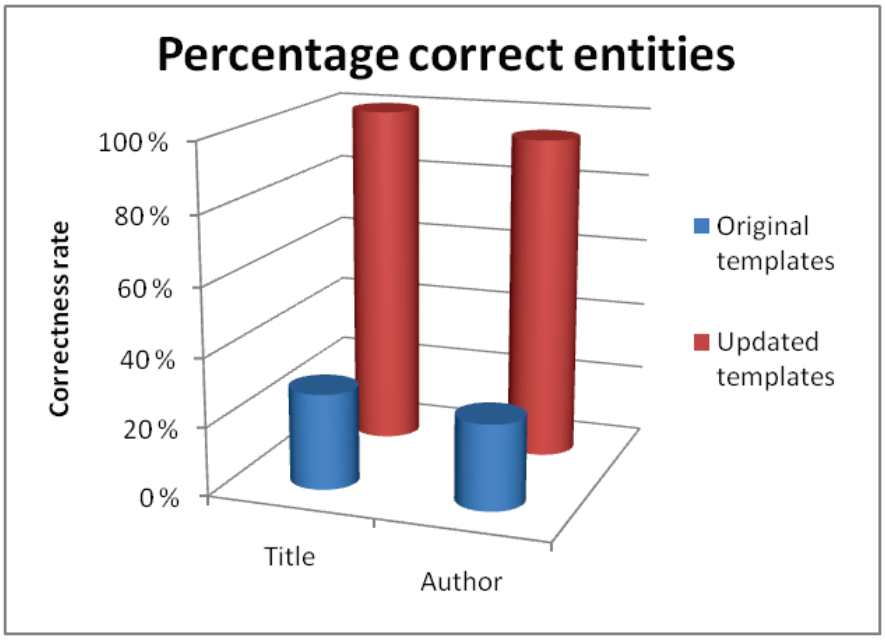


Figure 4: Results of the best algorithm on the different templates.

The limiting factor for the updated AMG efforts based on the Document Code is the human efforts. The AMG algorithm is only able to generate entities as good as the specified content of the document. This is in line with results from [8] where metadata were extracted using the Document Code on conference and journal papers.

6. CONCLUSION AND FUTURE WORK

AMG algorithms base their efforts on systematic and consistent properties of the documents at hand in order to generate quality metadata in accordance with pre-defined metadata schema(s). AMG algorithms need to find common structures in which to base their efforts, even if the dataset is not visually homogenous. The Document Code of each file format is a common structure which is shared by all documents based on the same file format version. Once access to the document is enabled, recognition of the most correct and most desirable document properties is essential in order to automatically generate high quality metadata. This paper has shown how such common characteristic can be utilized regardless of the visible characteristics of the document.

This paper has confirmed that AMG efforts based on visible characteristics are vulnerable for generating false entities if the dataset do not share a common set of visible characteristics. This paper has confirmed that the embedded document

metadata entities have extensive quality issues regarding pragmatics. The little awareness of the existence of embedded metadata among document authors contributes to this.

This paper has confirmed that AMG efforts based on the Document Code do not generate entities if the desired content is not present in the file. In the original Auditor dataset this resulted in no Author entities being generated and only a limited number of Title entities being generated. Some Meta tags are automatically generated by the used Word processor application. This includes Language Meta tags, enabling verification of the language of the intellectual content in high quality to sections as small as individual words.

This paper has shown how Meta tags can be included in document templates without changing the visible appearance of the original template. By including such Meta tags, document content can be uniquely identified and extracted regardless of the visible characteristics of the document. By doing this, this paper has enabled generation of metadata of a significantly higher quality in terms of pragmatics compared to the possibilities with the original documents. Still, when the user does not see or is not aware of the benefit of following the template, false content can be Meta tagged or desired content might not be Meta tagged at all. Both issues were discovered in the updated dataset.

There is a need for a close relation between the user, the used document template and the AMG algorithm in order to enable automatic generation of detailed and high quality metadata from common documents. Without such a relation, AMG algorithms cannot live up to their potential.

Future work should include: (1) Long term evaluation of how users use document templates when they are aware of how their documents are used for AMG efforts. (2) Research on the use of multi-linguistic documents in generating of semantic metadata using natural language approaches; (3) Analysis of the similarities between Latex templates in order to generate generic AMG algorithms based on the document code.

7. REFERENCES

[1] ADL. 2006. Sharable Content Object Reference Model (SCORM) 2004 3rd Edition Documentation Suite. http://www.adlnet.gov/downloads/AuthNotReqd.aspx?FileName=SCORM.2004.3ED.DocSuite.zip&ID=237

[2] Bird, K. and the Jorum Team. 2006. Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems. 31st March 2006, Version 1.0, Final", JISC and Intrallect July 2006

[3] Boguraev B. and Neff, M. 2000, Lexical Cohesion, Discourse Segmentation and Document Summarization. RIAO

[4] Bruce, T.R. and Hillmann, D.I. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions, In Metadata in Practice, D. Hillmann & E Westbrooks, eds., ISSN: 0-8389-0882-9

[5] Cardinaels, K., Meire M. and Duval E. 2005. Automating metadata generation: the simple indexing interface. Proc. of the 14th international conference on World Wide Web, Chiba, Japan, pp.548-556, ISBN: 1-59593-046-9

[6] Duval E. and Hodgins, W. 2004. Making metadata go away: Hiding everything but the benefits. Keynote address at DC-2004, Shanghai, China

[7] Edvardsen L.F.H. and Sølvberg, I.T. 2007. Metadata challenges in introducing the global IEEE Learning Object metadata (LOM) standard in a local environment. Proc.of WEBIST 2007, March 3-6, ISBN 978-972-8865-77-1, pp. 427-432, Springer

[8] Edvardsen, L.F.H. and Sølvberg, I.T. 2010. Could Automatic Metadata Generation be a digital solution for speedier and easier document publishing?. In review, 2010

[9] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T. and Trætteberg, H. 2009. Using the structural content of documents to automatically generate quality metadata. Proc. of Webist 2009, March 23-26, pp. 354-363, ISBN: 978-989-8111-83-8, ACM

[10] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T. and Trætteberg, H. 2009. Automatically generating high quality metadata by analyzing the document code of common file types. Proc. of JCDL 2009, June 15-19, ACM

[11] Edvardsen, L.F.H., Sølvberg, I.T., Aalberg, T. and Trætteberg, H. 2009. Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects. IEEE DEST 2009

[12] Flynn, P., Zhou, L., Maly, K., Zeil, S. and Zubair, M. 2007. Automated Template-Based Metadata Extraction Architecture. ICADL 2007. LNCS 4822, pp. 327–336

[13] Giuffrida, G., Shek, E.C. and Yang, J. 2000. Knowledge-Based Meta-data Extraction from PostScript Files. Digital Libraries, San Antonio, Tx, 2000, ACM 1-581 13-231-X/00/0006

[14] Google. 2010. Google. http://www.google.com

[15] Greenberg, J. 2004. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging (2004), 6(4): 59-82

[16] Greenberg, J., Spurgin, K., Crystal, A., Cronquist M. and Wilson, A. 2005. Final Report for the AMeGA (Automatic Metadata Generation Applications) Project. UNC School of information and library science

[17] Greenstone. 2007. Source only distribution. http://prdownloads.sourceforge.net/greenstone/gsdl-2.72-src.tar.gz (source code inspected)

[18] Hämäläinen, M., Whinston B.A. and Vishik, S. 1996. Electronic markets for learning: education brokerages on the Internet. Communications of the ACM archive, Volume 39, Issue 6 (June 1996), pp. 51 – 58, ISSN:0001-0782, ACM

[19] IEEE LTSC. 2005. IEEE P1484.12.3/D8, 2005-02-22 Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata. WG12: Related Materials. http://ltsc.ieee.org/wg12/files/IEEE_1484_12_03_d8_submitted.pdf

[20] Jenkins, C. and Inman, D. 2001. Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF. 0-7695-1022-1/01, 2001 IEEE

[21] Kawtrakul A. and Yingsaeree, C. 2005. A Unified Framework for Automatic Metadata Extraction from Electronic Document. Proc. of IADLC2005, 25-26 August 2005, pp. 71-77

[22] Li, H., Cao, Y., Xu, J., Hu, Y., Li, S. and Meyerzon, D. 2005. A new approach to intranet search based on information extraction. Proc. of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, pp. 460-468, ISBN:1-59593-140-6, ACM New York, NY, USA

[23] Lindland, O.I., Sindre, G. and Sølvberg, A. 1994. Understanding Quality in Conceptual Modeling. In IEEE Software, march 1994, Volume: 11, Issue: 2, pp. 42-49, ISSN: 0740-7459, DOI: 10.1109/52.268955

[24] Liu, Y., Bai, K., Mitra P. and Giles, C.L. 2007. TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries. JCDL'07, June 18–23, 2007, Vancouver, Canada, ACM 978-1-59593-644-8/07/0006

[25] LOMGen. 2006. LOMGen. http://www.cs.unb.ca/agentmatcher/LOMGen.html

[26] Meire, M., Ochoa X. and Duval, E. 2007. SAmgI: Automatic Meta-data Generation v2.0. Proc. of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007, pp. 1195-1204, Chesapeake, VA: AACE

[27] Open Archives Initiative. 2004. 2004 Protocol for Metadata Harvesting – v.2.0. http://www.openarchives.org/OAI/-openarchivesprotocol.html

[28] Scirus. 2010. Scirus – for scientific information. http://www.scirus.com

[29] Seymore, K., McCallum A. and Rosenfeld, R. 1999. Learning hidden Markov model structure for information extraction. Proc. of AAAI 99 Workshop on Machine Learning for Information Ex-traction, pp. 37-42

[30] Singh, A., Boley H. and Bhavsar, V.C. 2004. LOMGen: A Learning Object Metadata Generator Applied to Computer Science Terminology. National Research Council and University of New Brunswick, Learning Objects Summit Fredericton, NB, Canada, March 29-30, 2004

[31] Yahoo. 2010. Yahoo!. http://www.yahoo.com