USING THE STRUCTURAL CONTENT OF DOCUMENTS TO AUTOMATICALLY GENERATE QUALITY METADATA

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

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

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.

Edvardsen L., SÃ¿lvberg I., Aalberg T. and TrÃętteberg H.
USING THE STRUCTURAL CONTENT OF DOCUMENTS TO AUTOMATICALLY GENERATE QUALITY METADATA.
DOI: 10.5220/0001841003540363
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

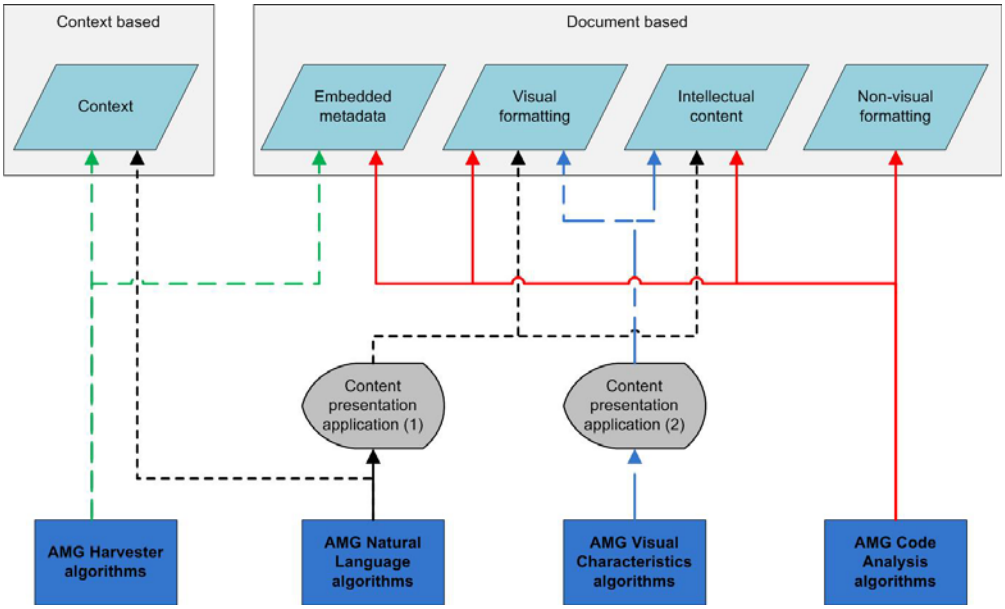

Figure 2: AMG content analysis algorithms and the data sources which they use.

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.

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.

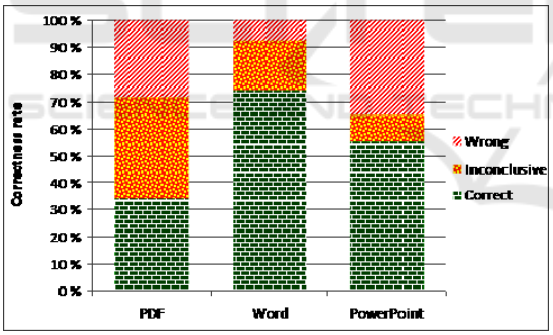

Figure 3: Documents created by LMS publisher

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.

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:

A Document Code Exclusively. Generates "Titles" elements based exclusively on the document code.

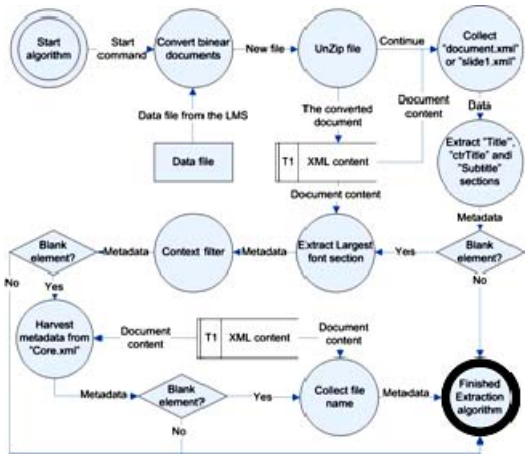

Figure 4: Logical structure of algorithm C.

A) 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.

B) 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 labeled 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 though 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.

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