Cape Peninsula University of Technology

31 Jan 2013

euDML Visibility to Google Free-form Searching – a Technical Report

A technical report on the accessibility of documents in the euDML to the Google crawler when using free-form searching

Author: Prof Melius Weideman, CPUT, Cape Town - WARC (Website Attribute Research Centre)

1. INTRODUCTION

The objective of the research behind this project was to do a pilot study on the visibility of a sample of 10 documents currently stored in the euDML digital library, to the Google search engine crawler. This digital library resides at: www.eudml.org . The rest of this Technical Report is arranged as follows:

2. ABSTRACT

A digital library user might not enter its webpages through the homepage menus – many are using standard search engines as opposed to academic databases to find academic information. Digital library owners should therefore not only consider the menus and/or information architecture as seen from the homepage. The aim of this research project was to determine to what extent the content of a mathematical digital library is indexed by Google. Free-form searching was considered, i.e. the use of a standard search engine via its search query box.

The euDML is a digital library (www.eudml.org) containing research documents on mathematical topics. An empirical pilot study was done to determine what the degree of visibility of a sample of these documents is to the Google search engine.

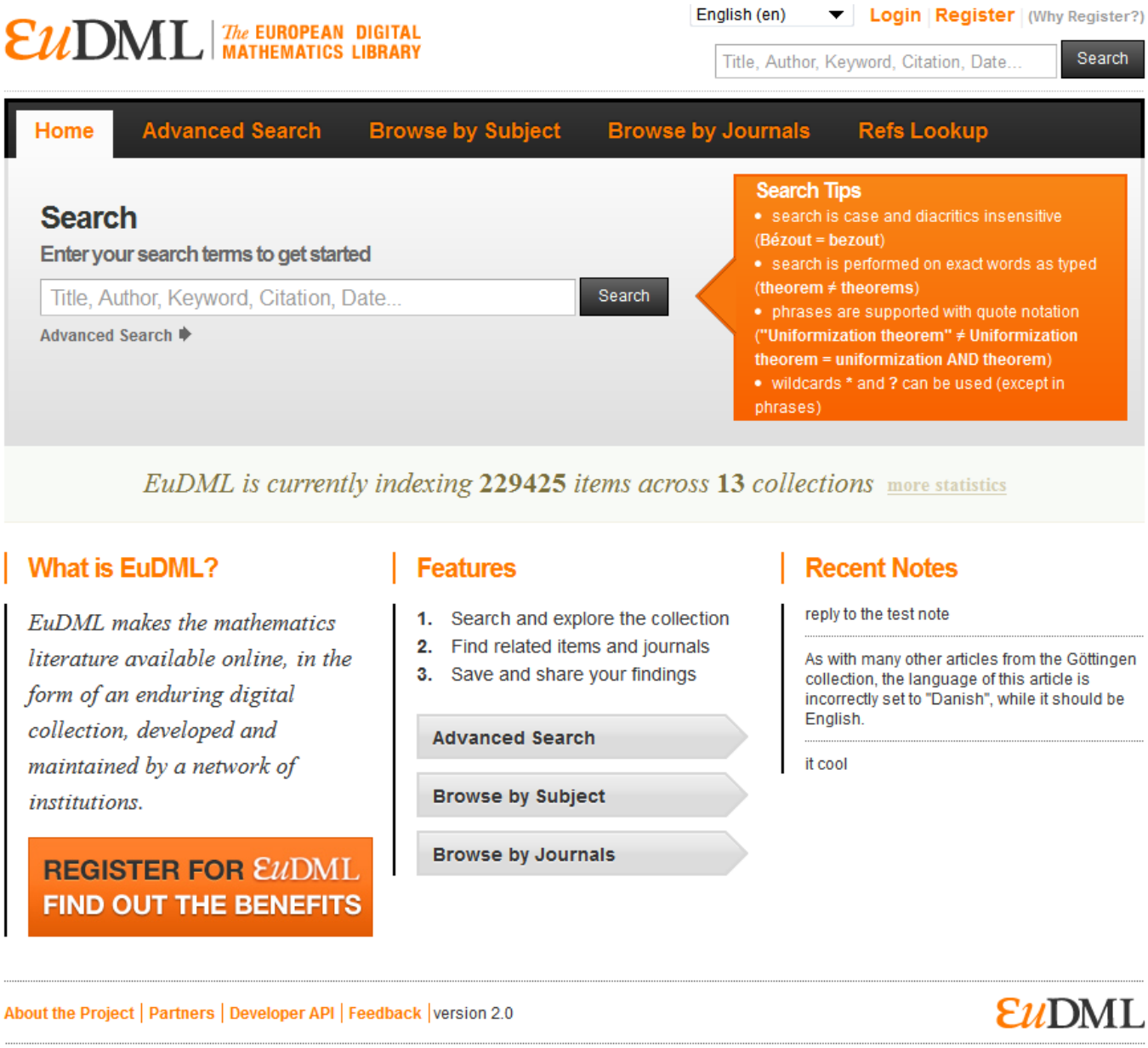
A random sample of indexed documents was supplied to the author by the staff behind euDML. Next, three search queries per document were generated, based on previous research done on the retrieval of PDF documents from academic digital libraries (http://www.academia.edu/263107/Empirical_Study_on_Crawler_Visibility_of_PDF_Documents_in_Digital_Libraries). These queries were based on the title, author surnames and body text of each article.

The first 20 results of each of the search engine result pages thus produced were inspected, and the presence/absence of the eudml.org domain or subdomains was noted.

Results prove that the full-text (stored as PDF documents) has virtually no visibility on Google, although some of the (presumably) metadata elements are indexed and ranks fairly highly on the result pages.

This research is based on the fundamentals of Website Visibility, as described in: Website Visibility: the theory and practice of improving rankings. Chandos Publishers, Oxford, UK. ISBN 1 84334 473 4. More details on this publication are at: http://www.web-visibility.co.za/website-visibility-abstracts-seo.htm

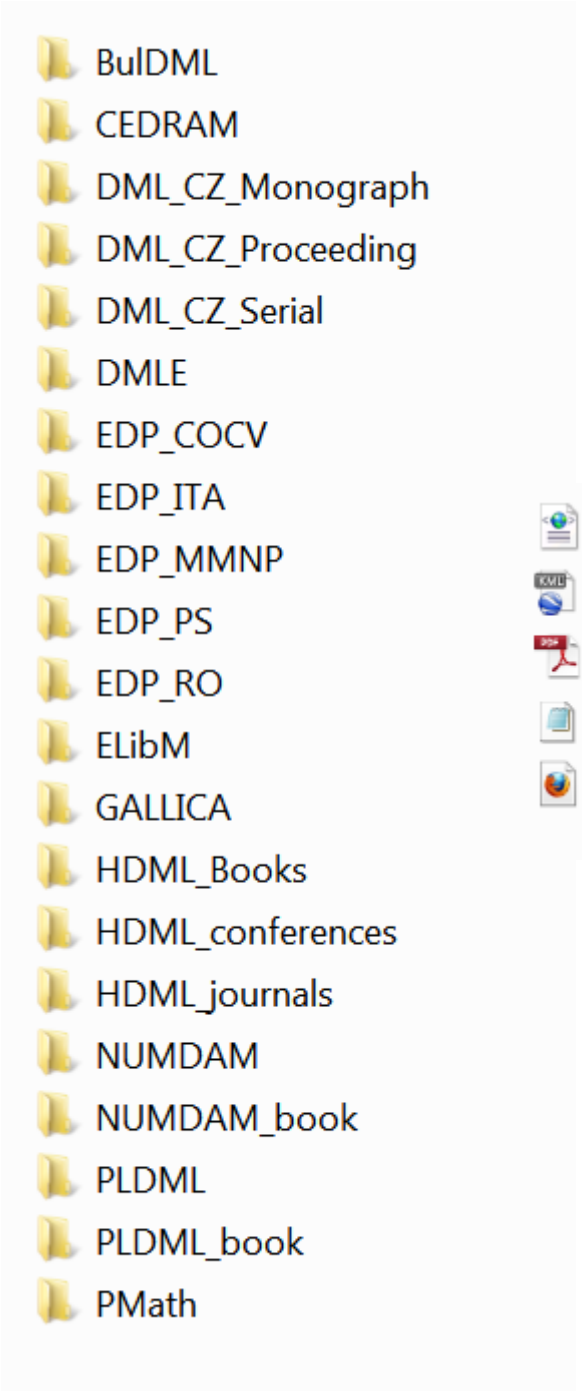
3. euDML HOMEPAGE

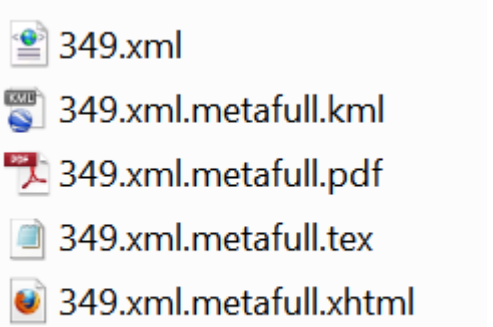


4. SAMPLE DOCUMENTS

The folders, as supplied to the author, are given in the image on the right. Each folder contained a number of documents, in most cases the PDF full-text of a given research output, as well as a metadata extract of the title, author, abstract, etc.

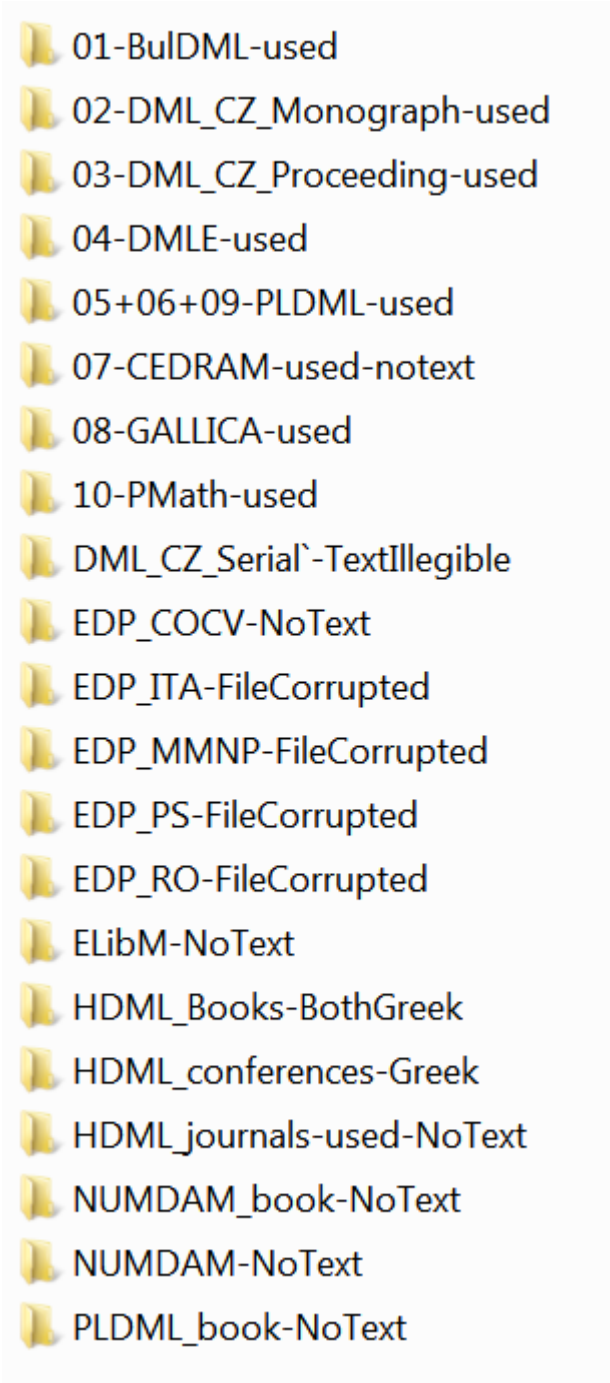
On the far right, the content of one sample folder is given. Every document had a unique ID number as part of the filename.





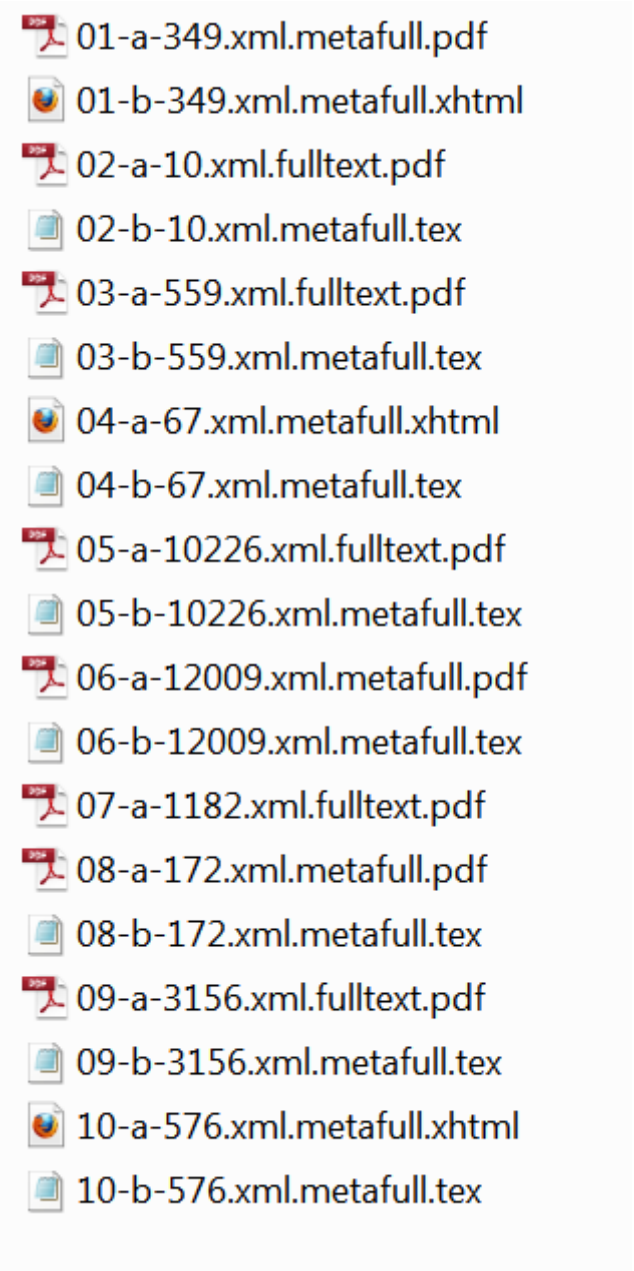
Some of the documents could not be used, for various reasons. In the image on the right, the 10 usable documents selected are those in the folders renamed to start with a 2-digit number.

The remaining folders have all been renamed with the reason for the document(s) being unusable given as the last part of the folder name.



The image on the right lists the names of the full-text and metadata of the 10 documents, which were used in this project. The unique digits after the “a-”or “b-”part of the filename indicate the ID of the chosen document.

The “a-” documents are the full-text PDF for each output, and the “b-” parts are the metadata. Document no. 7 did not have a metadata file, and the author manually generated the metadata from the PDF document.



5. SEARCH QUERIES GENERATED

In the results below, the main document language is specified, as well as some notes regarding the methodology used during this research. The text of Search Query 1 is also supplied. As described in previous research, SQ1 was generated by extracting up to five weight-carrying keywords, in sequence, from the document title.

Doc Lang	Doc no	Search Query 1
en	01	FRACTIONAL HELMHOLTZ EQUATIONS
en	02	No full-text or abstract could be found, doc in Czech language
en	03	FIXED POINT THEOREM INVERSE LIMIT
en	04	Banach Space
de	05	Über verallgemeinerte projektive Geometrie
en	06	Properties orthonormal Franklin
en	07	directions p-adic Hodge theory
fr	08	sur la formule de taylor
en	09	ZEROS PADe APPROXIMANTS CLASSES FUNCTIONS
en	10	HELSON SETS

	NOTES
1.	All OCR errors were fixed by comparing text to original, and applying basic rules of English.
2.	Some math formulae had to be converted and/or were converted by the copy process to simpler text.
3.	Commas were removed where body text would yield only 1 word, to increase span of search query.

In the results below, the text of Search Query 2 is given. As described in previous research, SQ2 was generated by concatenating the surnames of the authors of the document.

Doc Lang	Doc no	Search Query 2
en	01	Samuel Thomas
en	02	
en	03	MIODUSZEWSKI ROCHOWSKI
en	04	GODEFROY
de	05	Golab
en	06	CIESIELSKI
en	07	Kedlaya
fr	08	Liouville
en	09	KOVACHEVA
en	10	Miles

In the results below, the text of Search Query 3 is supplied. As described in previous research, SQ3 was generated by using the first sentence of the body text, up to but excluding the first punctuation mark. However, this rule was bent in some case – those which would produce only a single word as search query, which in turn would probably yield (unfairly so) no search results. Also, some mathematical formulae and/or notations were converted to equivalent text, for 2 reasons: a. the special characters could not be reproduced on a normal keyboard, b. the search engine query box would convert these characters in an unpredictable way, yielding invalid search results.

Doc Lang	Doc no	Search Query 3
en	01	The Helmholtz equation ∇2Ψ(x, y, z) +k2Ψ(x, y, z) =0 is named after Hermann Von Helmholtz
en	02	
en	03	A topological space X has the {\it fixed point property} (FPP) if for every continuous (single-valued) function f
en	04	We survey in these notes some recent progress on the understanding of the Banach space c_{0} an of its subspaces
de	05	eitdem F. Klein in seiner Arbeit "Über Liniengeometrie und metrische Geometrie α (1871) 1) die Beziehungen zwischen den projektiven und konformen Eigenschaften in der euklidischen Geometrie klargestellt hatte
en	06	The purpose of this paper is to present some properties of the orthonormal Franklin set and to indicate similarity of this system to another bases of the Banach space {\it C} $\langle$0,1$\rangle$ of continuous functions on $\langle$0,1$\rangle$
en	07	Throughout K will denote a finite extension of the field Qp of p-adic numbers
fr	08	Soit f(x) une fonction reelle de x
en	09	In the present paper, we deal with functions $f(z) :=\displaystyle \sum_{n=0}^{\infty}a_{n}z^{n}$ whose coefficients satisfy a special smoothness condition
en	10	A nonvoid compact subset e of t

6. SEARCH RESULTS

The results of the 27 searches done on the 10 documents are given in the image below. The 3 “G rank” columns list the ranking of each result found, successively for each query, colour-coded for easier interpretation. The first “G rank” column is the results from SQ1, etc.

A ranking result of “N” means “Not listed in the first 20 results”. This means that neither the domain www.eudml.org nor any of its subdomains were found on the first 2 result pages. It could be the case that it is present on results of the third and further down result pages. However, based on previous research, very few users view any results past the first, and even less past the second result page. For example, it is claimed that 67% of all clicks are on the FIRST TWO results on the first page, and that only 9% of readers request result pages after the third one supplied.

A ranking result where one or more digit appears, indicates that one of the relevant domains do appear in the first 20 results, in the position specified. A lower figure (i.e. higher ranking) is better. Two documents (no. 07 and 10) appear twice in the first 20 results, which is also a “better” result than just one figure.

An indication of “No Results” shows that the search did not produce even a single result for that query.

Doc Lang	Doc no	SQ1 G rank	SQ2 G rank	SQ3 G rank
en	01	N	N	No Results
en	02			
en	03	4	1	N
en	04	N	N	N
de	05	1	N	No Results
en	06	N	N	No Results
en	07	5 17	N	N
fr	08	N	N	N
en	09	2	N	N
en	10	1 17	N	N

7. RESULT INTERPRETATION – sample documents

The results from p10 are repeated here, to enable readers to follow the discussions with more ease. It is clear that the searches produced progressively worse results, from left to right. This is in line with previous research, which has proven that the "Keyword Extraction" method of query building (SQ1) is the most effective.

Since the 3 columns represent 3 different queries for the same document, and since the pattern of success/failure has been achieved before, it is suggested that these results not be interpreted on a per-column basis. In other words, the fact that the central column has only 1 green block, and the rightmost one none and 3 "No results", should not be considered as an indication of the document being present in Google's index or not.

Of more value is a per-row interpretation. Every row with at least 1 green block indicates a document with some exposure from the Google index. So, documents number 03, 05, 07, 09 and 10 are indexed and have rankings ranging from a top (position 1) to a reasonable (17) rank. The reason for document 03 being the only one listed for SQ2 (the "Surname" query), is probably the fact that this document had 2 quite uncommon surnames – arguably the most uncommon of the whole sample. This fact contributed to the high visibility of this document, as English text is over-represented on the Internet. Non-English keywords have a much higher chance of ranking well for any free-form search.

Doc Lang	Doc no	G rank	G rank	G rank
en	01	N	N	No Results
en	02			
en	03	4	1	N
en	04	N	N	N
de	05	1	N	No Results
en	06	N	N	No Results
en	07	5 17	N	N
fr	08	N	N	N
en	09	2	N	N
en	10	1 17	N	N

8. RESULT INTERPRETATION – other results

During the searching and result recording process, other results produced by Google were also briefly inspected. It was clear that a large number of documents (some of them PDF-type) of a mathematical nature did have exposure on Google, quite high up in the rankings. See some examples shown here.

Some of these platforms can be identified as being broadly “academic” by nature, similar to eudml.org - .edu, .org, researchgate, etc.

It was not clear (short of inspecting many of them – possibly worthy of another Technical Report) whether or not full-text of these documents was available, and whether or not this visibility was due to the exposure of metadata only.

It does prove however, that it is possible to expose PDF documents (possibly through text-based metadata, HTML file naming conventions, etc) to search engine crawlers, and to achieve some degree of website visibility.

[TeX] LaTeX Original
www.math.osu.edu/lectures/connes/zeta.tex
Lêerformaat: TeX/LaTeX - Bekyk as HTML
\smallskip These two problems will be solved in the **present paper**. **We** ... The second **N** osc (E) is a manifestation **of** the randomness **of** the actual ..... When Char (k) = **0**, **we** are **dealing** with an action **of** \Rb+* \sm \Rb on \Hcχ ..... In general (cf [Me]), E(**n**,s) is (-1)**n** times the **n**-th **coefficient of** the Taylor expansion at **z**=1 **of** ζs(**z**) ...

[PDF] Separation Theorems for Abstract Convex Structures
www.math.u-szeged.hu/.../kindler02.pdf - Vertaal hierdie bladsy
Lêerformaat: PDF/Adobe Acrobat - Kitskyk
property ∩{P : P ∈ Q} = ∅, and a **subset T** of S is called **compact** (in P) iff the ..... of all linear functionals f : **E** → R. Every **nonvoid** convex **subset** S of **E** can be ...

Personal Scheduler - Joint Mathematics Meetings
jointmathematicsmeetings.org/cgi-bin/mtg.../scheduler.pl?...
A new derivative concept for **set**-valued functions. Andreas H. Hamel ..... Shrinking and Boundedly complete frames for **Banach spaces**. Kevin Beanland, Virginia ...

Some new directions in p-adic Hodge theory | ResearchGate
www.researchgate.net/.../1764718_Some... - Vertaal hierdie bladsy
Throughout, K will denote a finite extension of the field Qp of p-adic numbers, and. GK = Gal(Qp/K) will denote the absolute Galois group of K. We will write Cp ...

9. SUMMARY, CONCLUSION and RECOMMENDATIONS

The results of this research were not entirely surprising, since it followed the pattern of previous research results.

In summary/conclusion:

a. None of the sample documents were indexed by body text – the specificity of the first sentences would have isolated these documents from the Google index had they been there. This proves that the PDF document body text was not indexed, and has little value for indexing by search engines. Again this is not really a surprise, as many of them appeared to be scanned copies in PDF format, making then unreadable as text by a crawler.

b. Half of the sample documents achieved top to good rankings on Google, when using the "Extracted Keyword" query method. This indicates that at least the titles of the sample documents were stored separately as text-based metadata, and this fact has been the major contributor to the high degree of visibility of these documents.

c. The "Surname Concatenation" method produces reliable search results only if the surnames contain one or more words which are highly uncommon in English, as for doc #03. Docs #5, 7, 9 and 10 are all indexed by Google, but they do not show up under the Surname search. All these surnames were single words (lower chance of success), and one of them is quite a common word (Miles). This confirms the basics of search query generation: single word and common word search queries have a low chance of success. Queries in excess of 2 words, and using uncommon words have a higher chance of success.

Recommendations:

The use of text-based metadata, the use of static webpages, a large volume of keyword-rich text and manual search engine submission are the keys to the achievement of high rankings in free-form Internet searching (see http://youtu.be/VAqxArE_6FE). Although the use of PDF documents has many advantages in digital libraries, their presence on the Web normally does not guarantee indexing and therefore they will probably not show up in search results on their own. Furthermore, the following recommendations should be considered:

1. Continue creating text-only metadata for every PDF stored in the digital library.

2. Focus on exposing these text-pages to search engines, as opposed to the PDF full-text documents.

3. Allocate manpower to manually submit text documents to search engines.

4. Consider other search engines apart from Google as being of value. Both Bing and Yahoo! have large followings outside Europe, and it is worth ensuring indexing by both these services, as well as by the top directory services. Furthermore, Baidu has a massive following (90% of all Internet users) in China. It has 78% of the search market, as opposed to Google's 18%. It reaches an audience double that of the USA and UK Google users combined – worth exposing content to.

5. The most important factor in increasing visibility is the quality and quantity of inlinks. Start the creation of a white hat Link Wheel structure, with cross-links between the different platforms. More detail can be provided – see point 7 below.

6. As a future project, the conversion and hosting of the full-text of euDML in text format must be considered. The body of writing of any high quality research document provides a rich harvest of honeycombed keyword text for search engine crawlers, and is a wasted opportunity if not used.

7. Finally, the author should be consulted for further research and recommendations on the visibility of the euDML content. There are many opportunities to increase its exposure, and expert advice should be used to design and oversee the implementation of a strategy for the advancement of this project in the world of search engines and users.

10. ACKNOWLEDGEMENTS

A final word of thanks must go to various people and organizations for playing a positive role in making this Technical Report possible.

- Prof Tomas Pitner from the Faculty of Informatics, Masaryk University from Brno, Czech Republic, for his support and collaboration on this project.

- Prof Petr Sojka (manager of the euDML project), also from FI, for his support and providing sample documents.

- Thanks also go to Masaryk University from Brno, Czech Republic, for hosting me for the 3 month Erasmus Mundus Scholarship, and to Erasmus Mundus for providing the opportunity for the Scholarship.

- To my employer, CPUT, (specifically Dr Nhlapo and Prof Staak) for releasing me to participate in this constructive Scholarship.

Prof Melius Weideman 31 January 2013