Insights from Over a Decade of Electronic Publishing Research

Fernando Loizides [a,1] and Sam A. M. Jones [a]
Emerging Interactive Technologies Lab
University of Wolverhampton, UK

Abstract. The work in this article presents findings from a text mining exercise of over a decade of research into electronic publication. We give readers insights into the past, present and possible future directions in a structured way, and further allowing them access to the extracted data in order to produce their own analysis and conclusion. We also produce our working methodology which can be replicated to produce systematic similar findings over the years as well as comparisons to other data souces.

Keywords. Electronic Publishing, Trends, Data Mining

1. Introduction and Methodology

The advent of the internet has brought about changes in the way that publishing takes place, both in terms of speed and cost. Research on electronic publishing has been reporting on current policies, stakeholder behavior and economic repercussions as well as other state of affairs. In order to understand the past, present, and perhaps even be able to hypothesize some of the near future of trending topics in electronic publishing, we present a text mining (Gupta 2009) exercise performed on over a decade of published material on electronic publications. In order to do this we take a sample from a leading electronic publication conference within Europe. The aim of the work is to identify to the reader areas in which research has steered over the years, report on likely trends and allow each reader to make their own inferences based on the data provided. Within the limitations such as scope, we aim to stimulate discussion and identify likely trends rather than absolute findings (a problem that would be more suitable to a big data analysis approach).

For our study we selected the corpus of the Electronic Publishing Conference from the years 2003 until 2015. A total of 462 full texts and 564 abstracts were extracted and converted into plain text format using a custom built scraper and format translator, built in the programming language python (*https://www.python.org*). The abstracts included in the full text documents but were also extracted as separate files in order for more specific analysis to be performed on these. We then performed a series of 'cleaning' activities on the data. This was required in order for a more effective lexical analysis to occur. The pseudocode for the cleaning process is as follows. (1) Remove common words (e.g. 'the' 'and') (2) Remove Specific Characters (e.g. '-' '.' '*' and punctuation

[1] Corresponding Author: fernando.loizides@wlv.ac.uk

marks) (3) Transform all text to lowercase (e.g. INFORMATION to information) (4) Remove Digits (it was deemed appropriate for our research that digits would not contribute to the findings) (4) Strip excess whitespace (5) Match Spelling (change all American to English or vice-versa) (6) Apply Porter Stemming Algorithm (Porter 1980) Stemming reduces words to their most basic state in order to identify similar words - e.g. 'visualize' and 'visualizing' would become 'visual'). The cleaning process took place using R (*https://www.r-project.org*). Once the cleaning process takes place, the documents were queried (again using R) as to the cumulative term frequency of each word across the documents (full texts and abstracts). This method is similar to the established 'analysis of co-occurring terms' (Buzydlowski et al 2002).

2. Findings and Discussion

The purpose of this section is to give the reader the high level findings from the data. The complete data findings can be found at (http://www.eitlab.net/wp-content/uploads/2016/03/elpub2016_DataMining.pdf). We focus on reporting a summary of the 'popular' areas in which electronic publishing has been reporting on. There are four subsections. The first presents the holistic findings from 2003 until 2015. The second section presents the findings from the last year of publication (2015) to report on the most recent trends but primarily to distinguish between the three year findings. The third section presents the cumulative findings of 2013, 2014, 2015 in order to distinguish a rate of change. The final section gives a more detailed year by year overview off the main findings from the last six years 2010-2015 to show longer term changes and suggest more stable areas.

2.1. 2003-2015

From the findings in the abstracts (See Table 1), we can see that almost no terms appear on more than 50% of the abstracts. Ignoring the words 'paper', 'use', 'publish', 'inform' and 'research' which are naturally occurring in an article, we highlight the words 'access' and 'develop'.

Table 1: Number of abstracts containing the same terms

Range	No of terms	Terms
324-524	0	
274-323	2	paper use
224-273	5	access develop inform publish research
174-223	5	base digit open present provid
124-173	16	also can content describ electron journal librari new project public result system technolog user web will

From the full texts (See Table 2) available since 2003, we are also able to highlight (occurring in over 75% of the documents) the terms 'access' and 'avail' (available, availability), while the documents also present several technology related terms such as 'http', 'technolog' and 'develop'.

Table 2: Number of documents (Full Texts) containing the same terms

Range	No of terms	Terms
412-462	4	can inform publish use
362-411	39	abstract access also avail base can confer data develop differ electron follow http import includ introduc keyword make need new one paper possibl present process provid refer relat research result system technolog time univers user web well will work
312-361	47	allow author case conclus content creat current digit discuss document exampl exist first form format group howev initi institut interest intern june level librari like manag mani may model number open order part proceed project public requir scienc search servic set specif support term text two way
262-311	83	activ addit anoth applic approach articl associ becom chang collect common communic communiti comput consid contain databas defin describ design direct distribut environ even experi field figur final find focus full function futur general high identifi implement increas integr issu journal knowledg languag larg link list made main mean object offer onlin particular point practice problem produc report repres resourc review see sever show sinc softwar sourc standard start still structur studi take technic three tool type version view wide within world year
212-261	119	abl academ accord ad address aim alreadi although among analysi appear appli archiv area aspect better book build call clear combin compar complet concept concern consist context contribut de descript detail effect eg elpub enabl end engin establish etc evalu expect fact featur found generat give given help human improv index indic individu internet involv last learn limit look mail major materi metadata method much must name nation natur necessari network non note now oper organ origin page perform person place potenti print product propos purpos qualiti question reason recent record relev repositori retriev right role scientif second select share signific similar social solut state step subject tabl th therefor thus toward tradit understand us valu various without word

2.2. 2015

The most recently available data come from the 2015 corpus and can be seen in Tables 3 (abstract) and Table 4 (Full Texts). Unsurprisingly we are able to distinguish the words ‘access’ and ‘open’ occurring in more than 50% of the abstracts, emphasizing the open access initiative that is highlighted and ever increasing in perceived importance by all stakeholders.

Table 3: Number of abstracts containing the same terms

Range	No of terms	Terms
20-26	0	
15-19	1	research
10-14	5	access also inform open paper
5-9	37	academ activ articl avail base benefit can current data develop digit experi find implement journal knowledg librari main new number particular practic present project provid public publish requir scienc scientif servic share studi use well within work

From the full texts for 2015, we are able to distinguish terms such as ‘model’, ‘project’ and ‘manag’, beyond the technological plethora of terms which are dominant.

Table 4: Number of documents (Full Text) containing the same terms

Range	No of terms	Terms
20-26	77	abstract access activ addit also associ author avail base can case consid correspond current data describ develop differ first follow form group howev http import includ increas inform institut interest introduc keyword knowledg level librari mail make model need new number one open order paper part particular practic present process project provid public publish refer relat requir research result review scienc servic share support system technolog term time univers use way web well will within work year
15-19	130	achiev aim allow alreadi among analysi anoth approach articl becom challeng chang collect common communic communiti compar conclus confer content context continu creat digit direct discuss distribut document enabl establish european even exampl exist expect experi field figur final format framework full futur general high https identifi impact implement improv initi integr intern issu journal key lead like link list made main major manag mani materi may mean measur method much must nation network now offer onlin organ particip place possibl potenti produc purpos qualiti question recent relev report repositori repres resourc right role scholar scientif search second see set sever sinc small social societi softwar solut sourc specif standard start state still structur studi subject suggest take technic text therefor third three topic toward two type user version wide

2.3. Three Years

Table 5 and 6 show the cumulative number of the terms occurring in abstracts and full texts respectively.

Table 5: Number of abstracts containing the same terms

Range	No of terms	Terms
46-57	0	
36-45	1	research
26-35	3	access open paper
16-25	12	base can data develop inform present provid public publish scienc use work

Following with the pattern of the previous observations in 2015, open access is reported in 50% or over of the documents. No other term (apart from the expected ‘research’ and ‘paper’) occur at this frequency.

Table 6: Number of documents (Full Text) containing the same terms

Range	No of terms	Terms
45-56	48	abstract access also author avail base can case current data develop differ follow http import includ inform institut introduc keyword level make mani need new one open paper possibl present process project provid public publish refer relat research result support system time univers use way web will work

35-44	98	activ addit aim allow analysi approach articl associ becom call chang collect communic communiti conclus consid content correspond creat describ digit discuss document even exampl exist experi figur final first focus form format futur group high howev identifi implement improv increas initi integr interest intern issu journal knowledg librari like link list made main major manag may mean method model nation number onlin order part particular practic recent report repositori require resourc review scienc search servic set share sinc social sourc start state structur studi take technic technolog term text therefor tool two type user well within year
25-34	158	abl academ accord achiev address alreadi among anoth appli applic archiv area basic best better challeng clear combin come common compar complet comput concept conduct confer consist contain context continu contribut core creation databas defin descript design detail direct distribut effect effort electron enabl end engin environ establish european expect extend fact factor field find found framework full function fund general give given global help human idea impact indic individu instanc interact investig involv lack languag larg lead learn least less long mail materi measur metadata might much must name natur network non now object offer often oper organ origin other output page perform place platform point polici potenti problem proceed produc propos purpos qualiti question read regard relev repres respons right role scholar scientif second see select sever show similar singl small societi softwar solut specif stage standard statist step still subject success suggest tabl third three thus topic toward valu various version view wide without world

The full texts of the 2013-2015 years provided more diversity in the results over the 50% level, with terms such as ‘social’, ‘knowledge’ and ‘management’ infiltrating and complementing the technological dominance and focus.

2.4. Year by Year (2010-2015)

We systematically studied each individual year between 2010 and 2015 (inclusive) and looked at the frequency of occurrences of words throughout the abstracts and full texts for that year. From the data we can see that there are 9518 terms appearing in either one, two, three or four papers. In 2014 the number of terms in this category is just 3159, an almost threefold decrease. One might be tempted to conjecture that there were more small clusters of similar papers in 2014 than in 2010 but if one considers the infamous birthday problem we are led to believe that there is less evidence for such a social explanation. The birthday problem (Wagner 2002) states that as the size of some collection of objects increases the number of ways of finding pairs within this collection increases much faster than the size of the collection. In 2014 there were 15 papers and in 2010 there were 35 (our dataset contains 32 of these 35 papers) so it is not surprising that there are many more terms in the one to four paper category – there are more ways of picking a pair or triplet or quartet of papers in 2010. In general it does look like there is some correlation between words appearing often in the abstracts and that same word appearing often in the papers (note again that we are looking across all documents here. A word appearing often in one particular abstract does tend to imply that it will occur often in the full text of that paper). It also appears that those words which do occur often in abstracts are words relating to general research, they do not necessarily appear to be particularly closely related to electronic publishing. When looking at the full text of the papers it appears that words which appear often are more likely to have something to do with publishing or electronic publishing in general. 2010

was a year with a large number of documents and so the numbers towards the bottom of the table tend to get very large. Interestingly, the phrase “http” appears in every single document from 2010. At a first glance it may seem that http appears as it is the standard prefix of URLs which often appear in the references section of academic papers. However, after some inspection it is clear that http appears in many of the papers outside of the references section and in some papers as a stand-alone term, although in the vast majority of cases it is still the prefix of a URL. There are seven words appearing in all of the papers from 2011. This number is 21 for 2012. In 2012 there were 30 accepted papers but our dataset contains only 19 of them. In 2011 there were 24 papers but the dataset also contains 19 documents. Whilst there are clearly many papers missing from the 2012 dataset it is interesting that the datasets have equal size and yet there are three times as many terms appearing in every paper in 2012. From the table we can see that the words which do not appear in every paper in 2011 are words such as “system”, “technology” and “digital” it might be reasonable to assume that the papers in 2012 were more focused on the systems and technological aspects of electronic publishing.

3. Conclusions and Future Work

In this work we present the findings from a text mining exercise of over 550 documents from over a decade of articles on electronic publication. We present the data in a structured form in order to give the readers, and different stakeholders, insight into the findings. Some high level comments identify interesting areas where terms correlate, without making any claim on statistical or predictive models for future directions for electronic publishing. The fragmented nature of the data set has a larger impact when trying to analyse trends from individual years than when trying to analyse trends over all the years. It is clear therefore that this section could be improved by improving the quality of the dataset. This could be done by obtaining access to more of the papers or by improving the quality of the text extraction from the PDF files. A possible direction for further work would be to extract some sort of contextual data from the files. Much of the analysis in this paper is based around the frequency of which words appear but if one were able to extract some contextual data it may be that more insightful inferences could be made. This is likely to be a difficult problem involving complicated NLP and text mining techniques and big data volume rendering techniques. We aim to continue the work on a much larger scale to achieve two goals. Firstly, verify if our data has further ecological validity and secondly, to identify further insights.

References

Buzydlowski, J.W., White, H.D. and Lin, X., 2002. Term co-occurrence analysis as an interface for digital libraries. In Visual interfaces to digital libraries (pp. 133-144). Springer Berlin Heidelberg.

Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp.60-76.

Porter, Martin F. "An algorithm for suffix stripping." Program 14, no. 3 (1980): 130-137.

Wagner, D., 2002. A generalized birthday problem. In Advances in cryptology—CRYPTO 2002 (pp. 288-304). Springer Berlin Heidelberg.