Metadata characteristics as predictors for editor selectivity in a current awareness service

Thomas Krichel*[a], Nisa Bakkalbasi[b]

[a] *Palmer School of Library and Information Science, Long Island University, 720 Northern Boulevard, Brookville, NY 11548, USA, http://openlib.org/home/krichel*
[b] *Kline Science Library, Yale University, 219 Prospect Street, PO Box 208111, New Haven, CT 06520-8111, USA*

Abstract

RePEc is a large digital library for the economics community. “NEP: New Economics Papers” is a current awareness service for recent additions to RePEc. The service is run by volunteer editors. They filter new additions to RePEc into subject-specific reports. The intended purpose of this current awareness service is to filter papers by subject matter without any judgment of their academic quality. We use binary logistic regression analysis to estimate the probability of a paper being included in any of the subject reports as a function of a range of observable variables. Our analysis suggests that, contrary to their own claims, editors use quality criteria. These include the reputation of the series as well as the reputation of the authors. Our findings suggest that a current awareness service can be used as a first step of a peer-review process.

1. Introduction

RePEc (Research Papers in Economics) is a large digital library for economics research. The roots of RePEc go back to 1993, when Thomas Krichel started to collect information about downloadable electronic working papers in economics. Working papers are accounts of recent research results before formal publication. Most economics departments in universities, as well as many other institutions that are involved in economics research—such as central banks and intergovernmental organizations—publish working papers. At the time of writing, over 430 data providers based at institutions that issue working papers contribute to RePEc. RePEc also contains article data from publishers such as Blackwell, Elsevier, and Taylor & Francis. All contributors provide classic bibliographic data about the papers that they publish, as well as links to the full text. At the time of writing, the dataset contains over 300,000 records. The data are harvested by service providers, who aggregate them to produce services for users who are interested in economics research. The RePEc web site at http://repec.org lists the service providers. All RePEc services are available to the public at no charge.

One of the RePEc services is NEP: New Economics Papers. NEP is a human-mediated current awareness service for RePEc's working paper data[1]. It primarily operates through electronic mail. It also has a web homepage at http://nep.repec.org. NEP has a simple, two-stage work flow. In the first stage, the general editor collects all new working paper data that have been submitted to RePEc in the previous week. She or he filters out records corresponding to

* Corresponding author. Tel: +1-516-299-2843, Fax: +1-516-299-4168.
E-mail addresses: krichel@openlib.org (Thomas Krichel), nisa.bakkalbasi@yale.edu (Nisa Bakkalbasi).

[1] Data about published articles is not included. In economics, the publishing delay is such that no published paper is new.

working papers that are new to RePEc, but are not new papers. Such records would typically come from new RePEc archives that add a whole back catalog of working papers to RePEc. The remaining records form a NEP report called nep-all. As its name suggests, nep-all contains all the new working papers in RePEc from the previous week. Each issue of nep-all is circulated to subject editors. This completes the first stage. In the second stage, the subject editors filter every nep-all issue they receive to contain papers in a certain subject area. When a new subject-specific issue has been created, it is circulated via e-mail to subscribers of the subject report. NEP is entirely run by volunteer subject editors from all corners of the globe. Most of them are PhD students or junior university faculty.

NEP was created by Thomas Krichel in 1998. Since that time NEP has grown in scale and scope. As RePEc has grown, so has the size of nep-all issues. This is scale growth. On the other hand, over time, more and more subject reports have been created. This is scope growth. At the time of this writing, there are close to 60 distinct subject reports in NEP. Over 13,000 unique e-mail addresses have subscribed to at least one NEP report. Over 45,000 new papers have been announced.

Chu and Krichel (2003) find that NEP is an interesting service model for digital libraries. Barrueco Cruz, Krichel, and Trinidad (2003) present a simple empirical assessment of the NEP service. One of the issues they look at is the subject coverage of NEP as a whole. They wonder if NEP covers all subjects that are found in RePEc. If it does, then we should observe that each working paper in a nep-all issue appears in at least one subject report. Empirically, this conjecture can be examined by looking at the ratio of papers in a nep-all issue that have been announced in at least one NEP subject report. This is called the coverage ratio of the nep-all issue. As more and more subject reports have been added, we expect the coverage ratio to improve over time, and in the longer run to reach 100%. Surprisingly, the data reported by Barrueco Cruz, Krichel and Trinidad (2003) suggest that the coverage ratio has not been improving as more reports have been added, and that certainly remains well below 100%.

In this paper, we are looking for explanations of this puzzle. We formally investigate subject editors' behavior. We want to find what makes a paper "announceable" in a NEP report. The remainder of our paper is organized as follows. In Section 2, we present some basic theory framework and describe our methodology for developing a predictive model. In Section 3, we discuss our data set. In Section 4 we present our findings. In Section 5 we develop our conclusions and suggest future work.

2. Theory and Methodology

We have two basic theories about editor behavior that aim to explain the static nature of the NEP coverage ratio. We call them the "target theory" and the "quality theory," respectively.

The target theory starts with the observation that the size of nep-all issues has been highly volatile in the short run, and has been steadily growing in the long run. The theory suggests that, when composing an issue of a subject report, the editors have an implicit issue size in mind. Therefore, if the size of nep-all is large, they will take a narrow interpretation of the subject matter of the report, i.e., they will be choosier as to what papers they include. Thus, the target theory claims that the observed long-run static nature of the coverage ratio comes from the simultaneous effect of scale and scope growth of NEP. Scale growth, all other effects being equal, will reduce the coverage ratio. Scope growth, all other effects being equal, will increase the coverage ratio. The long-run static coverage ratio is the result of both effects canceling each other out.

The quality theory suggests that subject editors filter for paper quality. There are two types of quality indicators. First, there is the descriptive quality of the record that describes a paper. Some papers are described poorly. They have a meaningless title, and/or no abstract. Second, there is the substantive quality of the paper itself. The paper may be written by authors whom nobody has ever heard of, and/or who are based at institutions with an unenviable research reputation. Whether it is substantive or descriptive, the quality of a paper is likely to be important when it comes to its inclusion in any NEP report.

A simple way to assess the target theory empirically is to see if the coverage ratio declines with the size of a nep-all issue. Barrueco Cruz, Krichel and Trinidad (2003) have a cross-sectional plot of coverage ratio versus size of nep-all. The shape of their plot suggests that this seems to be the case. However, they offer no rigorous statistical test. Even if inferential statistics were used, it would only look at one aspect of the selectivity issue. What we need is an overall model that combines a set of important variables to assess the probability of a working paper being "announced" in any report.

Statistically speaking, we conjecture that a percent of the variance in the response variable (i.e., the presence/absence of a paper in any report) can be accounted by predictor variables. If our conjecture turns out to be correct, we can produce a prediction equation that will allow us to predict the probability of a working paper being included in any NEP report. We believe that the most appropriate statistical method for analyzing this relationship is Binary Logistic Regression Analysis (BLRA). There are three reasons for our choice. First, the dependent variable is dichotomous. Each paper is either announced or not. Second, the independent variables are both quantitative and qualitative in nature. Last and most important, BLRA is a flexible technique. BLRA does not require any of the following assumptions commonly made for linear regression analysis to work:

- a linear relationship between the independent variables and the dependent variable
- homoscedasticity of the dependent variable for each level of independent variables
- a normally distributed dependent variable
- normally distributed error terms

According to Hosmer and Lemeshow (2000), BRLA originated in the epidemiological research. It is now heavily used in biomedical research. We expect that it will gain in popularity in Library and Information Sciences (LIS). We are not alone. In a review of statistical techniques in library and information sciences, Bensman (2001) suggests that, because of the highly skewed probability distributions observed in this discipline, researchers should look at the biomedical sciences for methodologies used to attack these issues.

3. Data Set

We have extracted data from historic e-mail message archives that contain NEP report issues. In addition, we have used the bibliographic records for the papers referred to in the report issues. The data go back to the inception of NEP in 1998 and contain 43,989 records, with each record corresponding to a paper that was appeared in nep-all. Our dependent variable is called ANNOUNCED. It takes the value 1 if the paper was announced in any subject report issue, and 0 if not.

We speculate that some observable variables influence the odds that a working paper makes into at least one subject report. The observable variables we can think of as relevant are

• SIZE	the size of the nep-all issue in which the paper appeared
• POSITION	the position of the paper in the nep-all issue
• TITLE	the length of a title of the paper
• ABSTRACT	the presence/absence of an abstract to the paper
• LANGUAGE	the language the paper is written in

- SERIES the size of the series where a paper appears in
- AUTHOR the prolificacy of the authors of the paper

SIZE is critical to assess the target theory. If target theory holds, we expect that an increase in SIZE reduces the odds of a paper being included. The variable POSITION can be motivated by an extension to target theory. Editors examine papers in nep-all one by one. An editor may find, as she moves to the bottom of the nep-all issue, that she has already fulfilled her implicit issue size target. Such an editor will be choosier with papers at the bottom of the nep-all issue than with papers at the top.

TITLE, ABSTRACT and LANGUAGE are the variables that are motivated by descriptive quality theory. The first, and many times the only thing that a subject editor looks in the bibliographic record is the title. If the title does not indicate the subject matter well, it is likely that the paper will be overlooked. We use the length of the title as a proxy for the amount of information it conveys. In a similar vein, ABSTRACT indicates the presence of an abstract[2]. Finally LANGUAGE stands for the language of the bibliographic record. We use the Lingua::Identify Perl module on the concatenated title, abstract and keywords to identify the language. If the abstract is missing the concatenated string is still rather short and the identification is unreliable. We create a dummy variable that takes the value 1 if the paper is in English, and 0 if it is in another language.

SERIES is the first variable that we use to asses substantive qualitative theory. Each working paper in RePEc appears as part of a series. We want to capture the reputation of a series. The bulk of RePEc working paper data come from departmental series, e.g. the Federal Reserve Bank of Minneapolis Research Department Working Paper Series. Usually, larger departments enjoy a higher reputation. They also issue more papers. There are also a number of large series with contributors based at many different departments. The National Bureau of Economic Research working paper series is an example. These large series enjoy a high reputation. Thus, it appears reasonable to use the size of a series as a proxy for its reputation.

AUTHOR is the second variable that we use to assess substantive qualitative theory. We use the prolificacy of authors as a proxy for their fame. There are at least three problems with measuring prolificacy. First, RePEc papers do not cover the entire economics discipline. Second, it is not easy to know if two similar author names represent the same person. Third, since co-authorship is frequent in economics, one needs to decide how to aggregate the prolificacy of individual authors. To deal with the second problem, RePEc runs an author registration service at http://authors.repec.org. This service collects records about the authors and the papers they have written. Authors contact the service to build their own electronic CVs out of RePEc data. Such registration is voluntary, of course. We do not have data available for all authors of all papers. At the time of writing roughly one third of all papers have at least one identified author. Therefore, we need to look for a measure that covers as many papers as possible. We use what we call the "lead author." For each paper, we take the number of papers in the RePEc database of the registered author with the largest number of papers. Still, due to a high number of unregistered authors, we end up with a significant number of records with missing values. After removing these records, we are left with 20,001 records. Since author registration and appearance of papers in reports are independent events, we conjecture that the removal of a large number of records introduces no sampling bias. To confirm, we analyze the descriptive statistics of both the original data set and the smaller data set. We find that the averages of the other independent variables stay approximately the same after the removal of roughly half the records. The size of the

[2] Initial intuition suggests constructing ABSTRACT as the number of characters in each abstract. However, proceeding in that way, we encounter a wide range of values [0, 11295], with the value 0 occurring very frequently. Conventional measures of central tendency and variance are meaningless in this context. Therefore we make ABSTRACT a categorical variable.

remaining data is still amply sufficient to conduct our analysis. Table 1 shows a few sample records from the data set before the removal of the records with missing values. In Table 1, HANDLE corresponds to the unique identifier for each record in the data set.

Table 1
Data set sample

HANDLE	ANNOUNCED	SIZE	POSITION	TITLE	ABS.	LANG.	SERIES	AUTHOR
jku:econwp:2001_05	0	230	.8996	88	1	en	47	NA
nbr:nberwo:9361	1	175	.7262	42	1	en	3654	NA
fip:fedfap:2002-02	1	433	.4646	54	1	en	96	105
wop:wobaiy:2957	0	803	.8691	66	0	en	12	NA
cbr:cbrwps:wp207	0	433	.2277	74	1	en	107	NA

All our calculations use the R language and environment. The R source code for this project, as well as our data set is available upon request.

4. Findings

4.1. Exploratory Data Analysis

First, we examine the data set to describe the main characteristics of each variable. We start with a frequency count of the response variable ANNOUNCED, reported in Table 2. It shows that nearly 73% of the 20,001 working papers are included in at least one subject-specific report.

Table 2
Frequencies of responses

ANNOUNCED	
0 = no	1 = yes
4,538	15,463

Table 3 displays the descriptive statistics for the quantitative predictor variables. It does not include the qualitative variables ABSTRACT and LANGUAGE. Neither does it include POSITION, of course.

Table 3
Descriptive statistics for quantitative predictor variables

	SIZE	TITLE	SERIES	AUTHOR
Minimum	3.0	3.0	1.0	1.0
1st quartile	146.0	48.0	66.0	12.0
Median	235.0	64.0	155.0	31.0
Mean	289.4	66.8	591.0	45.9
Standard deviation	195.0	29.4	1064.0	47.9
Coefficient of variation [a]	0.67	0.44	1.8	1.04
3rd quartile	379.0	82.0	498.0	64.0
Maximum	835.0	1945.0	3,654.0	385.0

For each quantitative predictor variable, we look at different measures of central tendency (i.e., mean and median) and dispersion (i.e., standard deviation, coefficient of variation, range). Two quantitative predictor variables, SIZE and TITLE, have their mean and median close to each other and their variation is small, implying a symmetrical distribution. For these two variables, the mean is an appropriate measure to determine their typical values for observation. Therefore, a typical working paper title contains 66 characters and the average nep-all size is 289.

For the variables AUTHOR and SERIES, there are significant differences between the mean and the median. The dispersion measures for those two variables indicate a high variation, due to the frequency of extreme values. To illustrate, there are many authors who have written 3 or 4 papers. There is only one author (Nobel laureate Joseph E. Stiglitz) with 385 papers. By the very fact that he appears as a prolific author on so many papers, he introduces an upward distortion for the average number of papers that an author has written. Therefore, for these two variables, the median qualifies as a more appropriate measure of central tendency than the mean. Thus, we conclude that an average lead author has written 31 papers, and that the average size of a series is 155.

4.2. Inferential Data Analysis

We use the "Design Library of Modeling Functions" in R to build the regression model. Prior to building the model, we perform a test to see whether the predictor variables are correlated to each other. Based on a Pearson correlation test among the 7 predictor variables, we conclude that there is no significant pair-wise correlation among any of the pairs of the predictor variables. Lack of correlation among the predictor variables increases our confidence that there is a higher likelihood for each predictor variable to contribute to the final prediction equation independently.

4.3. Fitting and testing the model

To create our model we use a collection of design variables (or dummy variables). For example, one of our independent variables SIZE is coded at three levels as "small," "medium," and "large." TITLE, and POSITION are also divided into three categories in the same manner.

Table 4
Coding the design variable for SIZE at three levels

	SIZE_1 = [179, 326)	SIZE_2 = [326, 835]
Small	0	0
Medium	1	0
Large	0	1

The results of the logistic regression analysis are shown in Table 5.

First, we need to assess how the overall model works. To that end, we perform the likelihood ratio test for the overall significance of the twelve coefficients for the independent variables. That is

H_0 : Coefficients for all variables equal 0

H_A :At least one coefficient is not equal to 0

Table 5
Estimated coefficients for multiple logistic regression model

			Coefficient	S.E.	Wald	P
Intercept			0.2401	0.08661	2.77	0.0056
SIZE_1	=	[179, 326)	- 0.2774	0.04317	-6.43	0.0000
SIZE_2	=	[326, 835]	- 0.4657	0.04262	-10.93	0.0000
TITLE_1	=	[55, 77)	0.1512	0.04108	3.68	0.0002
TITLE_2	=	[77, 1945]	0.2469	0.04222	5.85	0.0000
ABSTRACT			0.3874	0.04930	7.86	0.0000
LANGUAGE			0.7667	0.07850	9.77	0.0000
POSITION_1	=	[0.357, 0.704)	-0.0436	0.04238	-1.03	0.3039
POSITION _2	=	[0.704, 1.000]	0.0295	0.04226	0.70	0.4846

SERIES_1	=	[98, 231)	-0.1159	0.04109	-2.82	0.0048
SERIES_2	=	[231, 3654]	0.1958	0.04413	4.44	0.0000
AUTHOR			0.0001	0.00038	2.52	0.0118
			Model χ_2	d.f.	p-value	
			458.87	11	0	

The likelihood ratio test statistic takes the value 457.87. It is referred to as the model χ^2. This value leads us to reject the null hypothesis at virtually any significance level. We conclude that at least one of the twelve coefficients is different from zero and that, together, SIZE, TITLE, ABSTRACT, LANGUAGE, POSITION, SERIES, and AUTHOR, are significant predictors of ANNOUNCED. Next, we use the Wald statistics for the individual coefficients to test the significance of the variables in the model.

$$H_0 : \beta_i = 0$$
$$H_A : \beta i \neq 0$$

The Wald test statistics W are the ratio of each coefficient to its standard error.

$$W_i = \frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)}, \qquad \text{where } \beta_i \text{ is the coefficient for each predictor variable}$$

Based on the evidence contained in the data, at a significance level of $\alpha = 0.05$, we reject the null hypothesis for nine of the coefficients and conclude that SIZE, TITLE, ABSTRACT, LANGUAGE, SERIES, and AUTHOR are significant while POSITION is not significant. Table 5 contains the details.

Finally, we attempt to obtain the best fitting model with the least number of parameters. To that end, we run different models using different subsets of the predictor variables. After a thorough comparison of various models, we exclude POSITION from the final model. Let x be a vector representing values of the predictor variables. Then, the final logistic regression model, which gives us the estimated logistic probability, can be expressed as

$$(1) \qquad \text{P (ANNOUNCED =1}|\, x) = \frac{e^{\hat{g}(x)}}{1+e^{\hat{g}(x)}}$$

where the estimated logit is given by the following expression:

$$(2) \qquad \hat{g}(x) = 0.2401 - 0.2774{*}\text{SIZE\_1} - 0.4657{*}\text{ SIZE\_2} + 0.1512{*}\text{TITLE\_1} + 0.2469{*}\text{TITLE\_2} + 0.3874{*}\text{ABSTRACT} + 0.0001{*}\text{AUTHOR} + 0.7667{*}\text{LANGUAGE} - 0.1159{*}\text{SERIES\_1} + 0.1958{*}\text{SERIES\_2}$$

4.4 Interpreting the fitted logistic regression model

Equation (1) gives us a probability for the event occurring given all the values of the predictors. Equation (2) looks like a linear regression model as it is commonly understood. In such a linear regression equation, the coefficients are interpreted as the rate of change in the dependent variable associated with one-unit change in the respective independent variables. In the logistic regression model however, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable. Therefore the relationship between the independent and the dependent variable is less intuitive. One commonly-used measure of association is called the odds ratio, commonly abbreviated as OR. Roughly speaking, OR is a measure of the degree of association between each predictor variable and the outcome. It is obtained by transforming the estimated coefficients.

Table 6 contains the estimated OR values for our predictor variables. The interpretation of our two categorical independent variables ABSTRACT and LANGUAGE is straightforward. The OR for the ABSTRACT coefficient is 1.47, with a 95% confidence interval of [1.34, 1.62]. This suggests that in the presence of an abstract, a working paper is 1.47 times more likely to be announced than in the absence of an abstract. The estimated OR for LANGUAGE suggests that a working paper published in English is twice as likely to be announced then a working paper written in another language.

Table 6
Estimated odds ratios and 95% confidence intervals for predictor variables

	Odds Ratio (OR)	95% CI over OR
SIZE_1	1.32	[1.22, 1.44]
SIZE_2	0.83	[0.76, 0.90]
TITLE_1	1.16	[1.07, 1.26]
TITLE_2	1.28	[1.18, 1.39]
ABSTRACT	1.47	[1.34, 1.62]
LANG	2.15	[1.85, 2.51]
SERIES_1	1.11	[1.02, 1.20]
SERIES_2	1.37	[1.26, 1.49]
AUTHOR	1.05	[1.01, 1.09]

As we continue to examine the remaining variables, we notice that the OR reduces as the SIZE increases, indicating that working papers from large nep-all issues are less likely to be announced. When we examine TITLE we notice the odds for being announced increase as the title gets longer. Similarly, the OR estimates for SERIES suggest that the likelihood of a working paper being announced increases with the size of the series.

In summary, we presented statistical evidence that the predictive model based on the five predictor variables described in this section has potential to offer practical value in assessing the likelihood of a working paper submitted to NEP being included in at least one subject-specific report.

5. Conclusions and future work

In this paper we have considered the problem of choice between different sources of information. In conventional thinking about user needs, we think about users choosing information sources that fit closely their subject needs. We can think about this choice as horizontal choice between different sources that have essentially the same quality. In this paper, we have contrasted this horizontal choice with a vertical choice between different sources that are of different quality. Some users prefer a source that does not fit their subject need very closely if they think that the information is of high quality. This is a vertical choice between quality of sources. To empirically assess the strength of these components through editor interviews appears quite difficult. Subject editors, just like other information users, will find it difficult to evaluate their own choices.

Nevertheless before starting this work we inquired to the editors if they do any quality filtering. We used an informal discussion of the topic on the private email list used by the editors. Surprisingly enough, editors have a uniform view of vertical choice. They reject it. They claim that they perform their work independent of quality considerations. They assert that their only concern is to disseminate new working papers based on the subject matter. They specifically insist that NEP cannot be regarded as a vehicle for a preliminary peer review.

In this paper, we have set up and successfully tested a statistical model of editor selectivity in a current awareness service. The most important conclusion of our work is that the quality theory about editor behavior cannot be dismissed. That is, despite their assertion to the contrary, the subject editors appear to take into account the reputation of the series. Although we can only measure the reputation of the series in a very rough way, there still is statistical evidence of an editor bias. We also have shown that the prolificacy of the author has a significant impact on the announcement of their paper. As more authors register with RePEc we hope that we can show an even stronger impact of authorship.

The debate between horizontal and vertical choice has important implications for the running of NEP itself. If the target theory is correct, then opening additional specialized report categories should be considered as a way to improve the coverage of NEP. If the quality theory is correct, opening additional report categories will have little effect on coverage. Instead, it will have the adverse effect of making NEP more cumbersome to administer. This question of whether to open more reports or not has been one of the important motivations for the research conducted in this paper.

Since the quality theory can not be dismissed, NEP may be considered as a first stage in an alternative peer-review system. We could build such a system as sitting on top of NEP. It would be an extended service. In 2004, Thomas Krichel supervised work by a programmer to build a new interface for the construction of NEP report issues. This work is called the Altai project. A semi-formal specification is found at http://openlib.org/home/krichel/ work/altai.html. The new system, called “ernad” comprises a sorting stage. The editor has the opportunity to sort the papers in the subject report issue to move the “better” papers to the front. At the time of writing ernad has just gone live. Therefore we have no data on the uptake of this facility by the editors.

Ernad has been built with the target theory in mind. One important feature of ernad is that its workflow supports the idea of pre-sorting. Pre-sorting is supposed to sort the nep-all issue for a given subject in such a way that the editor finds the most likely candidates for inclusion at the top of the pre-sorted nep-all issue. Pre-sorting will use statistical learning techniques to discover the papers in the nep-all issue that are most likely to be included in the subject issue, given the historic behavior of the editor. At the time of writing, no specific learning procedure has been implemented. Nisa Bakkalbasi and Thomas Krichel are currently working on support vector machine techniques to implement a first test. The idea is that in the long run, NEP will be a system that combines machine learning with human guidance. When new subject reports are opened their editors usually are quite enthusiastic. They will carefully scan the nep-all issue. After a while, the novelty wears off and the editing job becomes quite boring. In the future, bored, established, old editors may be able to use the computer to sort the most likely papers to the front of the report. Thus they do not have to read the entire issue of nep-all. Scanning an entire nep-all issue is quite time-consuming. A diligent editor will roughly spent 10 minutes to scan 100 papers. With nep-all issues of 600+ papers this is a significant task. With a part of the horizontal choice being made by the computer, the editors can spend more energy on sorting the report issue by bringing the best papers to the top of the report. This is a subtle, long-run push from a current awareness to a peer-review orientation of NEP.

References

Barrueco Cruz, J. M., Krichel, T., & Trinidad, J. C. (2003). *Organizing current awareness in a large digital library*. Presented at the 2003 Conference on Users in Electronic Information Environments in Espoo, Finland, September 8-9, 2003, http://openlib.org/home/ krichel/papers/espoo.pdf.

Bensman, S. J. (2000). Probability distributions in Library and Information Science: A historical and practitioner viewpoint. *Journal of the American Society for Information Science and Technology,* 51(9), 816-833.

Bookstein, A. (2001). Implications of ambiguity for scientometric measurement. *Journal of the American Society for Information Science and Technology,* 52(1): 74-79.

Chu, H. & Krichel T. (2003). NEP: Current awareness service of the RePEc Digital Library. *D-Lib Magazine, 9*(12). http://www.dlib.org/ dlib/december03/chu/12chu.html

Hosmer, A. W. & Lemeshow S. (2000). *Applied logistic regression*. New York, USA: John Wiley & Sons

Maindonald, J. & Braun, J. (2003). *Data analysis and graphics using R – an example-based approach.* Cambridge, UK: Cambridge University Press.