The citation advantage of open access articles

Loughborough University

Thesis Access Form

Copy No……...……Location…….
……...…

Author……Michael Norris…….

Title……The citation advantage of open access articles
……

Status of access OPEN / RESTRICTED / CONFIDENTIAL

Moratorium period:……years,ending……../……200……
……..…..

Conditions of access proved by (CAPITALS):…….Professor Cliff McKnight ..…
Director of Research

(Signature)……...……
……...

Department of Information Science
……...……

Author's Declaration: I agree the following conditions:

OPEN access work shall be made available (in the University and externally) and reproduced as necessary at the discretion of the University Librarian or Head of Department. It may also be copied by the British Library in microfilm or other form for supply to requesting libraries or individuals, subject to an indication of intended use for non-publishing purposes in the following form, placed on the copy and on any covering document or label.
The statement itself shall apply to ALL copies:

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
Restricted/confidential work: All access and any photocopying shall be strictly subject to written permission from the University Head of Department and any external sponsor, if any.

Author's signature…….Date……...……
…...……...

users declaration: for signature during any Moratorium period (Not Open work): I undertake to uphold the above conditions:			
Date	Name (CAPITALS)	Signature	Address

The citation advantage of open access articles

Michael Norris

Doctoral Thesis

Submitted in partial fulfilment of the requirements

for the award of PhD of Loughborough University

May 2008

Certificate of Originality

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements, footnotes or references, and that neither the thesis or the original work contained therein has been submitted to this or any other institution for a higher degree.

……Signed

……Date

Abstract

Four subjects, ecology, applied mathematics, sociology and economics, were selected to assess whether there is a citation advantage between journal articles that have an open access (OA) version on the Internet compared to those articles that are exclusively toll access (TA). In two rounds of data collection, citations were counted using the *Web of Science* and the OA status of articles was determined by using the search tools *OAIster*, *OpenDOAR*, *Google* and *Google Scholar*. In the first round a purposive sample of 4633 articles for the four subjects from high impact journals were examined, 2280 (49%) were OA and had a mean citation count of 9.04, whereas the mean for TA articles was 5.76. There was a clear citation advantage for those articles that were OA as opposed to those that were TA. This advantage, however, varied between disciplines, with sociology having the highest citation advantage but the lowest number of OA articles from the sample taken and ecology having the highest individual citation count for OA articles but the smallest citation advantage. Tests of correlation between OA status and a number of variables were generally found to be weak or inconsistent but some associations were significant. *Google* and *Google Scholar* were more successful at finding OA articles on the Internet than were *OAIster* or *OpenDOAR*. The country of origin of the citing authors for applied maths was found in order to assess whether those authors from poorer countries cited OA articles more frequently than TA articles. While cited to citing article ratios from lower income countries favoured OA articles, overall percentages gave mixed results.

The data from the second round confirmed the result for sociology. The second sample for ecology was randomly taken from 82 journals and exhibited a greater OA advantage. For economics, a second purposive sample of articles from 21 mid-range impact journals was taken and also exhibited a greater OA advantage. In an attempt to establish the cause of any citation advantage, logistic regression was used to try to determine whether the bibliographic characteristics of the articles from both rounds could be used to predict OA status. Results from this were generally inconclusive.

Keywords: Open Access: Citation Advantage: Causation: Logistic Regression.

Acknowledgements

I wish to thank my two supervisors Professor Charles Oppenheim and Dr Fytton Rowland for their constructive and intelligent guidance. Thanks go to my Director of Research Professor Cliff McKnight.

Thanks to all those staff in the general office for their kindness and help and all the many staff in the department who over the years have been party to my many endeavours. Special thanks go to Dr Richard Gadsden who gave key statistical guidance and advice.

Two of my fellow students deserve particular thanks, Fadi Qutaishat for his fellowship and Layla Hasan for her immense good humour.

Finally, my thanks go to Loughborough University for funding and nurturing me.

Contents

ABSTRACT I
ACKNOWLEDGEMENTS II
CONTENTS III
LIST OF TABLES IX
LIST OF FIGURES XI
GLOSSARY OF TERMS AND ACRONYMS XIII
CHAPTER 1 INTRODUCTION 1
1.1. Background 1
1.2. Aims and objectives 7
CHAPTER 2 LITERATURE REVIEW: SCHOLARLY COMMUNICATION 10
2.1. Introduction 11
2.2. Scholarly communication 13
2.3. The origins of the academic journal 13
2.4. The growing number of journals and scientists 15
2.5. The growth of commercial journal publishers 17
2.6. Commercialisation of journal publishing 18
2.7. Journal costs 20
2.8. The serials crisis 22
2.9. Bundling and big deals 25
2.10. Online access and measuring usage 27
2.11. Peer review 28
CHAPTER 3 LITERATURE REVIEW: OPEN ACCESS 32
3.1. OA Movement – definition 33
3.2. Drivers to OA 34
3.3. Pre and postprints 35
3.4. Models of OA 35
3.5. The green route 36
3.6. The gold route 37
3.7. Author pays 39
3.8. Delayed access 41
3.9. Summary 41

3.10. Interoperability and the Open Archives Initiative 42
3.11. Repositories 43
3.12. Distributed institutional repositories 44
3.13. Self-archiving to repositories 45
3.14. Mandatory self-archiving 48
3.15. Mandatory self-archiving and publishers 49
CHAPTER 4 LITERATURE REVIEW: BIBLIOMETRIC TECHNIQUES 51
4.1. Bibliometrics 52
4.2. History of bibliometrics 53
4.3. Laws of bibliometrics 54
4.4. Lotka's Law 55
4.5. Bradford's Law 56
4.6. Zipf's Law 57
4.7. Cautions and convergence of the laws 57
4.8. Citation analysis 58
4.9. Journal coverage and impact factor 60
4.10. The meaning of citations 61
4.11. Why people cite 61
4.12. Issues with citation analysis 64
4.13. Validity of citation studies 65
4.14. OA and research impact 67
4.15. Research impact in discrete disciplines 67
4.16. Research impact within particular journals 70
4.17. Research impact across multiple disciplines 73
4.18. Causation 75
CHAPTER 5 METHODS 82
5.1. Introduction 83
5.2. Background: Information Science and bibliometrics 83
5.3. Background: brief history of the scientific method 85
5.4. Methodology: the research strategy 87
5.5. Methodology: understanding correlation 88
5.6. Methodology: deductive theory 90
5.7. Methodology: research design 93
5.8. Methodology: literature review 94

5.9. Methodology: pilot studies 94
5.10. Methodology: previous research strategies 95
5.11. Justification: database and search tools selection 97
5.12. Justification: selection of OA/TA article search tools 100
5.13. Justification: subject selection 101
5.14. Methods adopted 103
5.15. Methods adopted: literature review 103
5.16. Methods adopted: pilot studies 104
5.17. Methods adopted: data collection 105
5.18. Methods adopted: general data collection 106
5.19. Methods adopted: particular objectives 107
5.20. Integrity: data management 113
5.21. Integrity: data analysis 113
5.22. Integrity: replication, reliability and validity 114
5.23. Integrity: pitfalls to be avoided 115
5.24. Integrity: limitations of the study 115
5.25. Integrity: strengths of the study 117
CHAPTER 6 RESULTS: FIRST ROUND DATA COLLECTION 118
6.1. Introduction 119
6.2. Data overview 119
6.3. Distribution of citation counts 120
6.4. Self-citation counts 126
6.5. Author frequency and OA/TA status 131
6.6. Correlations 137
6.7. Search engine success 139
6.8. Impact factor 142
6.9. Within journal comparisons 145
6.10. Distribution of citations within subjects 148
CHAPTER 7 RESULTS: SECOND ROUND DATA COLLECTION 151
7.1. Introduction 152
7.2. Objective 2 152
7.3. Distribution of citation counts 152
7.4. Self-citation counts 155
7.5. Author frequency and OA/TA status 158

7.6. Correlations 161
7.7. Search engine success.....161
7.8. Impact factor.....164
7.9. Within journal comparisons 164
7.10. Distribution of citations within subjects.....165
7.11. Objective 3.....167
7.12. Data overview.....167
7.13. Distribution of citation counts.....167
7.14. Self-citation counts.....171
7.15. Author frequency and OA/TA status.....174
7.16. Correlations 177
7.17. Search engine success.....178
7.18. Impact factor.....180
7.19. Within journal comparisons 181
7.20. Citation distribution.....182
7.21. Objective 4.....185
7.22. Data overview.....185
7.23. Distribution of citation counts.....185
7.24. Self-citation counts.....188
7.25. Author frequency and OA/TA status.....191
7.26. Correlations 194
7.27. Search engine success.....195
7.28. Impact factor.....198
7.29. Within journal comparisons 198
7.30. Distribution of citations within subjects.....199
7.31. Objective 5.....201
7.32. Countries of origin.....202
7.33. Logistic regression analysis – collective subjects 211
7.34. Logistic regression analysis – individual subjects.....218
7.35. Logistic regression analysis – second round individual subjects 219
CHAPTER 8 DISCUSSION 221
8.1. Introduction 222
8.1.1. Background, aims and objectives 222
8.1.2. Research questions 223
8.1.3. Areas of interest 224

8.2. Open access - overall results 224
8.2.1. Levels of open access 225
8.3. Citations, self-citations, correlations and associations 229
8.3.1. Citation advantage 229
8.3.2. Citation distribution 231
8.3.3. Self citations 232
8.3.4. Within journal comparisons 233
8.4. Correlation, associations and OA status 234
8.4.1. Authorship levels, subjects and OA status 234
8.4.2. Country of origin 236
8.4.3. Impact factor and OA 237
8.5. Discovery of OA articles. 238
8.6. Causation – in general. 242
8.6.1. Possible causation: ecology 243
8.6.2. Possible causation: OA and poorer countries 243
8.6.3. Causation in collective subjects 246
8.6.4. Causation in individual subjects 247
8.7. Conclusion 248
CHAPTER 9 CONCLUSION AND RECOMMENDATIONS 250
9.1. Introduction 251
9.2. Main findings 251
9.3. Contribution of the study 251
9.3.1. Citation advantage 252
9.3.2. Levels and distribution of OA articles 252
9.3.3. OA article characteristics 252
9.3.4. Search tool success 252
9.3.5. Causation 253
9.4. Implications 253
9.4.1. Government 254
9.4.2. Institutions 254
9.4.3. Funders 255
9.4.4. Information suppliers 255
9.4.5. Academics 257
9.5. Recommendations 257
9.5.1. Government 257
9.5.2. Institutions 257
9.5.3. General information suppliers 258
9.5.4. Google Scholar 258
9.5.5. Metadata harvesters 258
9.5.6. Academics 259
9.6. Limitations of the study and recommendations for further research 259
9.7. Recommendations for further research 259
BIBLIOGRAPHY 262
Appendix A - Pilot studies 290
Appendix B - Journal titles 301
Appendix C - Outlier details 308
Appendix D - Journal titles 309
Appendix E - Journal titles 311

Appendix F - Individual journal citation advantage 312
Appendix G - Logistic regression variables 319
Appendix H - SPSS output 320
Appendix I – Publications 326

List of Tables

Table 3.1 Author pays (acceptance) fees for publishing a single article.....42
Table 4.1 Illustration of Lotka's law – author productivity (Oppenheim [n.d]). 56
Table 5.1 Qualitative and quantitative research strategies 88
Table 5.2 The research design used in this research. 93
Table 5.3 Work that has identified a citation advantage for OA articles. 96
Table 5.4 Subject coverage and orientation from Moed (2005, pp.129-130) 102
Table 5.5 Principal search terms and combinations 104
Table 6.1 Gross citation counts 123
Table 6.2 Citation count net of author and journal self-citations 123
Table 6.3 Key citation data for the four subjects.....125
Table 6.4 Mean number of authors per article 132
Table 6.5 Author counts by article frequency 132
Table 6.6 OA/TA article counts by country and subject. 136
Table 6.7 Correlation by subject for author/number of citations 138
Table 6.8 Correlation by subject for impact factor and numbers of authors.....139
Table 6.9 Break down of OA hits by subject and search tool 140
Table 6.10 Search tool success by subject and region.....142
Table 6.11. Impact factor outlier details.....143
Table 6.12 Distribution of all citations by percentage article count.....148
Table 6.13 Distribution of other author only citations by percentage article count 149
Table 6.14 Journal titles selected for citation distribution 149
Table 6.15 Distribution of all citations by percentage article count at journal level.....150
Table 6.16 Distribution of author only citations by percentage article count 150
Table 7.1 Search tool success by subject and region.....163
Table 7.2 Distribution of citations by percentage article count.....166
Table 7.3 Distribution of citations by percentage article count at journal level.....166
Table 7.4 Search tool success by subject and region.....179
Table 7.5 Distribution of citations by percentage article count.....182
Table 7.6 Search tool success by subject and region.....197
Table 7.7 Distribution of citations by percentage article count.....200
Table 7.8 Distribution of citations by percentage article count at journal level.....200
Table 7.9 Cited articles by region and OA status 202
Table 7.10 Citing articles by region and OA status.....203
Table 7.11 Frequency of citation.....204
Table 7.12 Ratio of citing to cited articles by income group 205
Table 7.13 Cited to citing articles by income group 206
Table 7.14 Regional citation match by author country 207
Table 7.15 Citing to cited articles by region - % by column.....208
Table 7.16 Citing to cited articles by region - % by row 209
Table 7.17 Citing country to cited income group.....210
Table 7.18 Classification table for regression model OA/NOA.....215
Table 7.19 Iteration history 215
Table 7.20 Omnibus results.....216
Table 7.21 Hosmer and Lemeshow results.....216
Table 7.22 Final classification table at step eight.....216
Table 7.23 Model if term removed.....216

Table 7.24 Variables left in equation after eight steps 217
Table 7.25 Logistic regression results at subject level – first round data.....218
Table 7.26 Logistic regression results at subject level – second round data.....219

List of Figures

Figure 3.1 Repository numbers (ROAR 2008).....45
Figure 4.1 Age distribution of citations (Moed 2007).....79
Figure 4.2 Citation profile (Kurtz & Henneken 2007).....80
Figure 5.1 The field of Information Science 84
Figure 5.2 Possible causal relationships.....90
Figure 6.1 Proportion of OA/TA articles by subject 119
Figure 6.2 Distribution of all citations 120
Figure 6.3 OA and TA citation distribution 121
Figure 6.4 Citation distribution by subject 122
Figure 6.5 Frequency of non-cited articles by subjects.....122
Figure 6.6 Boxplot of the distribution of all citations 124
Figure 6.7 Other author citations.....125
Figure 6.8 Distribution of self-citations by OA/TA status 126
Figure 6.9 All OA and TA self citations 127
Figure 6.10 Total self-citation count by subject.....128
Figure 6.11 Boxplot of self-citations.....128
Figure 6.12 Breakdown of OA citations by subject 129
Figure 6.13 Breakdown of TA citations by subject.....129
Figure 6.14 Percentage citations by their citation category 130
Figure 6.15 OA/TA scatterplot of self-citations to other author citations.....131
Figure 6.16 Articles by author count and OA/TA status.....133
Figure 6.17 Author count by OA/TA status and subject 134
Figure 6.18 Author count by region and OA status.....135
Figure 6.19 Number of OA/TA articles by region 136
Figure 6.20 Article count and OA/TA status by region 137
Figure 6.21 OA/TA author citation scatterplots 137
Figure 6.22 Search tool success rate.....140
Figure 6.23 OA article hits by region and search tool.....141
Figure 6.24 Boxplot of impact factor scores by subject.....143
Figure 6.25 Impact factor by OA/TA status 144
Figure 6.26 Impact factor by OA/TA status and subject.....145
Figure 6.27 OA/TA article split for ecology 146
Figure 6.28 OA/TA article split for economics.....146
Figure 6.29 OA/TA article split for applied maths.....147
Figure 6.30 OA/TA article split for sociology 147
Figure 7.1 Distribution of all citations 153
Figure 7.2 OA and TA citation distribution article 153
Figure 7.3 Boxplot of the distribution of citations 154
Figure 7.4 Boxplot for other author citations 155
Figure 7.5 Distribution of self-citations by OA/TA status 156
Figure 7.6 All OA and TA self citations 156
Figure 7.7 Breakdown of TA/OA citations 156
Figure 7.8 Articles by their citation category.....157
Figure 7.9 OA/TA scatterplot of self-citations to other author citations.....158
Figure 7.10 Articles by author count and OA/TA status.....158
Figure 7.11 Articles by author count and region.....159

Figure 7.12 Author count by region and OA status....160
Figure 7.13 Number of OA/TA articles by region....160
Figure 7.14 OA/TA author citation scatterplots....161
Figure 7.15 Search tool success rate....162
Figure 7.16 OA article hits by region and search tool....163
Figure 7.17 Impact factor by OA/TA status....164
Figure 7.18 OA/TA article spilt by journal title....165
Figure 7.19 Distribution of all citations....168
Figure 7.20 OA and TA citation distribution....168
Figure 7.21 Boxplot of the distribution of all citations....169
Figure 7.22 Boxplot of other author citations....170
Figure 7.23 Distribution of the OA advantage by mean citation count....170
Figure 7.24 Distribution of self-citations by OA/TA status....171
Figure 7.25 OA and TA citation distribution article....172
Figure 7.26 Boxplot of self-citations....172
Figure 7.27 Breakdown of TA/OA citations....173
Figure 7.28 Articles by their citation category....173
Figure 7.29 OA/TA scatterplot of self-citations to other author citations....174
Figure 7.30 Articles by author count and OA/TA status.....175
Figure 7.31 Articles by author count and region....175
Figure 7.32 Author count by region and OA status....176
Figure 7.33 Number of OA/TA articles by region....177
Figure 7.34 OA/TA author citation scatterplots....177
Figure 7.35 Search tool success rate....178
Figure 7.36 OA article hits by region and search tool....179
Figure 7.37 Impact factor by OA/TA status....180
Figure 7.38 Impact factor by OA status....181
Figure 7.39 Gross mean citation profile by regions....183
Figure 7.40 Lower impact journals by mean citation count....184
Figure 7.41 Distribution of all citations....185
Figure 7.42 OA and TA citation distribution article....186
Figure 7.43 Boxplot of the distribution of all citations....187
Figure 7.44 Boxplot of other author citations....187
Figure 7.45 Distribution of self-citations by OA/TA status....188
Figure 7.46 All OA and TA self citations....189
Figure 7.47 Boxplot of self-citations....189
Figure 7.48 Breakdown of TA/OA citations....190
Figure 7.49 Articles by their citation category....190
Figure 7.50 OA/TA scatterplot of self-citations to other author citations....191
Figure 7.51 Articles by author count and OA/TA status.....192
Figure 7.52 Articles by author count and region....193
Figure 7.53 Author count by region and OA status....193
Figure 7.54 Number of OA/TA articles by region....194
Figure 7.55 OA/TA author citation scatterplots....195
Figure 7.56 Search tool success rate....196
Figure 7.57 OA article hits by region and search tool....197
Figure 7.58 Impact factor by OA/TA status....198
Figure 7.59 OA/TA article split for economics....199
Figure 7.60 Percentage of citations by OA status....211

Glossary of Terms and Acronyms

arXiv	Pre and postprint disciplinary archive set up in 1991, specialises in astronomy, physics, computer sciences and mathematics. [http://arxiv.org/].
ALPSP	Association of Learned and Professional Society Publishers. An association representing the interests of not-for-profit publishers.
AGORA	Access to Global Online Research in Agriculture. Set up by the UN with journal publishers to allow access, by developing countries, to electronic journals at minimal cost.
ARL	Association of Research Libraries. Principally a North American coalition of university libraries.
ASC	Author self-citations. Where the authors cite themselves.
Authors Pays	Model of payment where authors or their institutions pay the cost of publishing an article in an academic journal.
Big Deal	Publisher's delivery model for the bulk purchase by libraries of journals more usually in electronic format.
Bibliometrics	The study of patterns of authorship, publication and literature by the use of various statistical techniques.
Bundling	A practice of journal publishers, supplying in bulk journals in paper and electronic form to libraries.
Citation	A reference to a document which usually appears as a footnote, endnote or in a bibliography.
Citation Impact	The impact of an article/document as counted by the number of citations it has received.
DOAJ	Directory of Open Access Journals. An online database of journals that are Open Access. [http://www.doaj.org/]
Eprint	Term used to describe both pre and postprint articles.
Gold Route	Model of journal publishing where readers can access the contents without charge, authors or author institutions usually pay for publication.
Green Route	Option where authors can self-archive their eprints to an electronic repository.

HEFCE	Higher Education Funding Council for England. Body which distributes funds to higher education in England. Scotland, Wales and N. Ireland have their own councils.
HINARI	Health Internetwork Access to Research Initiative. Body set up by the WHO with publishers to allow access, by developing countries, to electronic journals at minimal cost.
Impact Factor	Measure devised by Thomson ISI which is used to rank peer reviewed journals by their average citation count.
Institutional Archive	An electronic archive into which members of an institution can self-archive their work. Interchangeable term with institutional repository
Institutional. Repository	Interchangeable term for Institutional Archive.
ISI	Institute for Scientific Information, commercial company which maintains the *Web of Science* – a database of academic journals and their citation records.
JASC	Journal author self-citations. Where the authors have cited themselves from an article, which appeared in the same journal in which they are writing
JISC	Joint Information Systems Committee. Organisation funded by Higher Education Funding Councils to provide guidance on the use of information computing systems to support learning and teaching.
JSC	Journal self-citation. Where any writer cites an article from the journal, in which they are writing themselves.
LISU	Library and Information Statistics Unit. Department within Loughborough University which collects and publishers statistics related to libraries and information science in general.
Metadata	Used here to describe the basic bibliographic data of an article – author, article title, journal title etc.
NIH	National Institutes of Health. American governmental body responsible for funding research in medical related fields of research.
OA	Open Access. Term used to describe journals, their articles or other documents which are freely available to all, to copy, distribute and use without impediment apart from acknowledging their author/source.
OAI	Open Archives Initiative. An initiative that develops and promotes interoperability standards that aims to facilitate the efficient dissemination of eprints and other electronic content.

OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting. Framework and guidance for the design of interoperable systems that can be used by others to design software that can be used to harvest metadata from electronic archives.
OAIster	OAIster is a union catalogue of digital resources. It provides access to digital resources by "harvesting" their descriptive metadata from OAI compliant repositories. [http://www.oaister.org/].
OC	Other citations. Citations made by authors that are unrelated in any way to the original article they are citing, other than the act of citation itself.
OpenDOAR	In part a search tool that allows users to search repositories worldwide using *Google*'s indexes of repositories that have been 'crawled' by their web crawler. [http://www.opendoar.org/]
Peer Review	A process where (usually) journal articles are critically reviewed by academic peers and if necessary revised prior to their acceptance for publishing in an academic journal.
Postprint	A version of a peer reviewed article, which has not usually been copy edited but has been self-archived by the author to a repository usually after the fully edited version has been published in an academic journal. This is known as the Green Route to self-archiving.
Preprint	An article written by an author usually self-archived to a repository.
RAE	Research Assessment Exercise. A peer review evaluation exercise carried out to measure the quality of a university's research output.
RCUK	Research Councils UK. Coordinating body made up of the representatives of the eight funding councils which distribute a significant proportion of government research funding.
Research Impact	Impact made by research as measured by citation counts.
RoMEO	Rights Metadata for Open Archiving. A project funded by the Joint Information Systems Committee to investigate the rights issues surrounding the self-archiving of research in the UK.
ROAR	Registry of Open Access Repositories. Database of repositories and their associated deposit statistics.
Self-Archiving	The act of placing an electronic document into an electronic archive which can be accessed by anyone.
SHERPA	Securing a Hybrid Environment for Research Preservation and Access. Is developing open-access institutional repositories in a number of research universities.

SPARC	Scholarly Publishing and Academic Resources Coalition. Mostly North American coalition of academic and research libraries that is trying to reduce the impact of the cost of purchasing scholarly work.
STM	Scientific technical and medical. A group of publications, which are particular to this area of knowledge and supported by a range of publishers who are often specialists in these disciplines.
TA	Toll access. Term used to describe journals or their articles, which can only be viewed by payment of a subscription.
Wellcome Trust	A UK charitable organisation which funds medical research.
WoS	Web of Science. Database of resources linked primarily to the counting and analysis of citations in scholarly publications.

Chapter 1 Introduction

1.1. Background

In their research and in their interpretation of studies over the last forty years Tenopir and King (2000, p.4) have shown that reading, and particularly the reading of academic journals, improves the efficiency and quality of subsequent research. They also consider that "over the last three and one-half centuries, scientific scholarly journals have become the principal means of publishing scientific information" (2000, p.5). In publishing their findings, academics wish primarily to make their work available to their peers. Whilst clearly not the only medium for publishing research, publications in peer reviewed academic journals are critical for researchers in most disciplines, academic or otherwise, who want to gain recognition for their work. The process of scholarly communication is an integral part of the research cycle; it allows for the dissemination and discussion of research findings, permitting them, if credible, to become the building blocks for further research (Marks 2001, p.91). This process of recognition not only enhances the academic's reputation, but makes it more likely that they can expect from the community in which they work an acknowledgement of their contribution with tokens of preferment, status and esteem, including citations (Thompson 2005, pp.45-46).

Scholarly communication through the academic journal typically requires the author to give, in its entirety and without financial reward, their work to their publisher (Harnad 2004a, p.64). In this process, authors will probably have also surrendered their intellectual property rights to their work as well. In this bargain, the author hopes in return to be read as widely as possible by others in their field of work. The status, rank and impact of particular academic journals in which authors aspire to be published, is readily recognised by those in any particular discipline (de Vries 2001, pp.250-251). Authors are helped in this process of identifying the rank, however contentious, of leading journals in their discipline by standardised metrics, which rank them (Page, Campbell & Meadows 1997, p.137). Commonly, this is achieved by counting the number

of times that other academic authors have cited the articles appearing in a particular journal (Garfield 1979, p.24). From these citation counts a simple but powerful ranking tool has been devised which ranks journals by their 'impact factor' (IF) within a particular subject. The IF "is calculated by dividing the number of current year citations by the source items published in that journal during the previous two years" (The ISI Impact Factor...[n.d.]). Crudely, the greater the number of citations that a journal receives for its articles, the more likely it is to be more highly regarded and ranked by its readership and those authors who aspire to publish in it (Page, Campbell & Meadows 1997, p.137).

Until the advent of the Internet and the World Wide Web, the model of scholarly communication was, as originally founded, still based on paper. The Internet, however, has made it possible, for the journal article in particular, to be found, accessed and delivered electronically from locations remote from their paper based source (Tenopir & King 2000, pp.38-39). It is also possible now for scholars to communicate electronically their research, their views and thoughts on their work and the work of others by a variety of means (Borgman 2000, p.4).

Despite this electronic freedom, the traditional model of journal publishing effectively remains intact, with publishers controlling access to the majority of journal titles (Cox 2003, p.13). Such access requires a payment by the individual reader or an institution, which makes payments on their behalf. These journals may be described as toll access (TA) (Harnad 2004a, pp.64-65). Journal publishers range in size, business model and ownership, from commercial publishers driven by profit to learned societies who publish on a not-for-profit basis but who use any surpluses generated to promote the aims of their society (Cameron 2001, p.247). Noticeably, in the last few decades, large commercial publishers have managed the journal market to their advantage by increasing journal costs beyond the costs of inflation and library budgets (Thompson 2005, p.100). This process has made the cost of holding journals, however they are accessed, a considerable and growing part of an institution's budget. The market in academic books has suffered noticeably in this process, with library expenditure being skewed in favour of maintaining subscriptions to serials (Thompson 2005, pp.103-107). The Association of Research Libraries (2008 pp.12-16) statistics for 2005-06 show that serials costs have been

increasing much faster than inflation with serial expenditure having risen by 321% over the last twenty years, compared to only 82% for monographs in the same period.

The rise in the cost of serials has fuelled, in part, a movement towards the possibility of accessing the results of academic research without charge (Guedon 2001, pp.15-17). Since the mid-1990s, there has been a movement to allow open access (OA) (Willinsky 2006, p.35.) in particular, to peer-reviewed scholarly articles. Although there are many definitions of OA, the one given by Suber (2006), gives a straightforward account of the OA movement, describing it as:

> Putting peer-reviewed scientific and scholarly literature on the internet. Making it available free of charge and free of most copyright and licensing restrictions. Removing the barriers to serious research.

The proponents of OA contend that publicly funded research should be accessible to anyone who wishes to read it, and at no cost. Currently, for the majority of academic journals this is not the case; readers or their institutions have to pay a significant and growing cost to access this material (House of Commons Science and Technology Committee 2004, p.29). Not only is removing the financial barriers to access an issue, but it is possible that the potential impact of research is being lost, in that it cannot be easily shared and built upon. Attempts have been made to quantify this loss of research impact in financial terms. For the UK, Harnad (2005a) estimates that £1.5bn pa may be being lost because of restricted access. This estimate has however, been subject to scrutiny and criticism, not the least of which is that the basis of his calculation relies on the estimated value of a citation (Harnad 2005b). Harnad uses the work of Diamond ([n.d.]), who estimated the value of a single citation as being between $50-1300 (US) in 1986, depending on its discipline and number of citations. However, as Garfield ([n.d.]) points out in an essay which includes analysis of Diamond's work, "Readers should be cautious in drawing certain conclusions from Diamond's article. Diamond is not saying that every additional citation is worth "X" amount of dollars".

To institutional readers, whose institutions have paid the necessary journal subscription fees, journals and their contents are effectively free at the point of use. Quite clearly, to the institution and most likely the library which is buying the subscription to these

journals, the costs are high and continue to rise (Tenopir & King 2000, pp.273-274). The pattern and models of access to the scientific literature varies, depending on how it is funded (House of Commons Science and Technology Committee 2004, pp.21-28). The committee reports (2004, p.25), for example, that for the general public "making information available only for a high fee at the point of access has the most severe repercussions for one particular group of end-users; patients". For those who have access to the Internet, the promise of freely accessing academic journal content is becoming a possibility that is free of charge at the point of use. Allowing anyone to freely access academic journal content is the principal aim of the OA movement. Journal publishing does, however, require an adequate income stream to remain viable and successful. New financial models of funding are being used which make it possible for readers to access some journals without charge; Willinksy (2006, pp.212-213) lists ten variants of OA:

Home page	University department maintains home pages for faculty members on which they place their papers and make them freely available.
E-print archive	Institution or academic subject area underwrites the hosting and maintenance of repository software enabling members to self-archive published and unpublished materials.
Author fee	Author fees support immediate and complete access to open access journals (or in some cases, to the individual articles for which fees were paid), with institutional and national membership available to cover author fees.
Subsidised	Subsidy from scholarly society, institution and/or government/foundation enables immediate and complete access to open access journal.
Dual mode	Subscription fees are collected for print edition and used to sustain both print edition and online open access.
Delayed	Subscription fees are collected for print edition and immediate access to online edition, with open access provided to content after a period of time (e.g., six to twelve months).
Partial	Open access is provided to a small selection of articles in each issue - serving as a marketing tool – whereas access to the rest of the issue requires subscription.

Per Capita	Open access is offered to scholars and students in developing countries as a charitable contribution, with expense limited to registering institutions in an access management system.
Cooperative	Member institutions (e.g. libraries, scholarly association) contribute to support of open access journals and development of publishing resources.

Typically, these models have found a variety of sources of funds to support their journal access. The most significant of these are where the author or author's institution pays for publication, often from research budgets or where the journals are able to draw on substantial gifts from donors (Willinsky 2006, pp.211-216). There are also a growing number of electronic, interoperable institutional or disciplinary repositories into which academics, with the agreement of their publishers, can self-archive their work. Self-archiving is defined as the process of:

> …deposit[ing] a digital document in a publicly accessible website, preferably an OAI-compliant Eprint Archive. Depositing involves a simple web interface where the depositer copy/pastes in the "metadata" (date, author-name, title, journal-name, etc.) and then attaches the full-text document. (Self-Archiving FAQ 2005).

In this arrangement, there can be various embargo periods specified by the publisher before such archiving or free access can take place (SHERPA…[n.d.]). Ironically, given that there is some evidence to support open access to journal articles by virtue of their greater impact, and that the majority of publishers are now prepared to allow authors to self archive their work to these repositories, authors have been slow to use this resource (Swan & Brown 2005, pp.26-27). They suggest that authors would self-archive their work if asked to do so, but few have done so voluntarily. This reluctance to self-archive has given rise to calls for publicly funded research to be deposited routinely on a mandatory basis (Swan *et al.* 2005, p.33). This is a highly contentious issue and has drawn publishers, funding bodies and those both for and against OA into prolonged debate. Latterly, however, there have been moves for example in the UK for government funded research councils and the National Institutes of Health in the USA which have mandated the archiving of the output from publicly funded research.

An academic is frequently judged, in part at least, on the quality of their published research. The greater the impact of that research as counted by, for example, the number of citations it receives, the better, it is believed, is the quality of the work (van Leeuwen *et al.* 2003, pp.262-263). Receiving many citations for academic research generally correlates very strongly with academic success; a simple analysis of Nobel laureates and their citation counts by Ashton & Oppenheim (1978), Garfield (1992) and Opthof (1997) gives significant credibility to the idea that the two are linked. It seems logical, then, that self-archiving academic research into OA archives where the work is likely to be read and cited more often should, in turn, lead to greater impact. Some research to support this has come from large-scale comparisons between work that is freely available and that which is not (Harnad & Brody 2004). The metric used to make this comparison is citation counts. If it can be shown that research output that is OA receives more citations than closed access research, then a convincing argument can be made to persuade researchers to archive their work or use OA journals. There have been a small number of other studies, (see for example Antelman 2004, Harnad & Brody 2004, Eysenbach 2006) which have supported this hypothesis, but these have generally been limited to a small number of disciplines or have been confined to specific disciplinary archives. The arXiv archive that houses preprints from high-energy physics, mathematics and computer science has been used extensively to demonstrate this citation advantage and explore the causes of this advantage.

At first glance, it seems elementary to argue that articles that are OA will receive more citations than those that are not, they should be easily found, read and cited by a larger number of academics and thus have a greater impact in their discipline. This view has been challenged by journal publishers (see Craig *et al*. 2007) who argue it is also possibly damaging, especially if authors readily made their work OA in an attempt to garner a citation advantage. In fact Kurtz *et al*. (2005), Moed (2007) and Kurtz & Henneken (2007) have, using material from the arXiv, begun to demonstrate that suggesting any citation advantage derived from an article being OA is far too simplistic. They rather think that making articles freely available prior to journal publication combined with the quality of its authors and/or its content are the reason for any citation advantage.

1.2. Aims and objectives

The primary aim of the research was to examine whether OA articles, as measured by citation counts, have greater impact than their toll access counterparts and also to consider what might be the causes of any such citation advantage.

The following research questions were posed:

- Do OA articles receive more citations than comparable TA articles and thus have greater research impact?
- Is there an identifiable causal link between any citation advantage and an article's OA status?

These gave rise to the following null hypotheses:

- That there is no significant difference in the rates of citation between open access and toll access articles in favour of open access articles, in a group of subjects over a fixed period.
- That there are no causal links between the open access status of an article and the collective bibliographic details associated with that status.

To test the validity of these hypotheses, five objectives were set:

Objective 1: Determine the OA citation advantage or otherwise by examining the citation counts of high impact journal articles from four discrete disciplines.

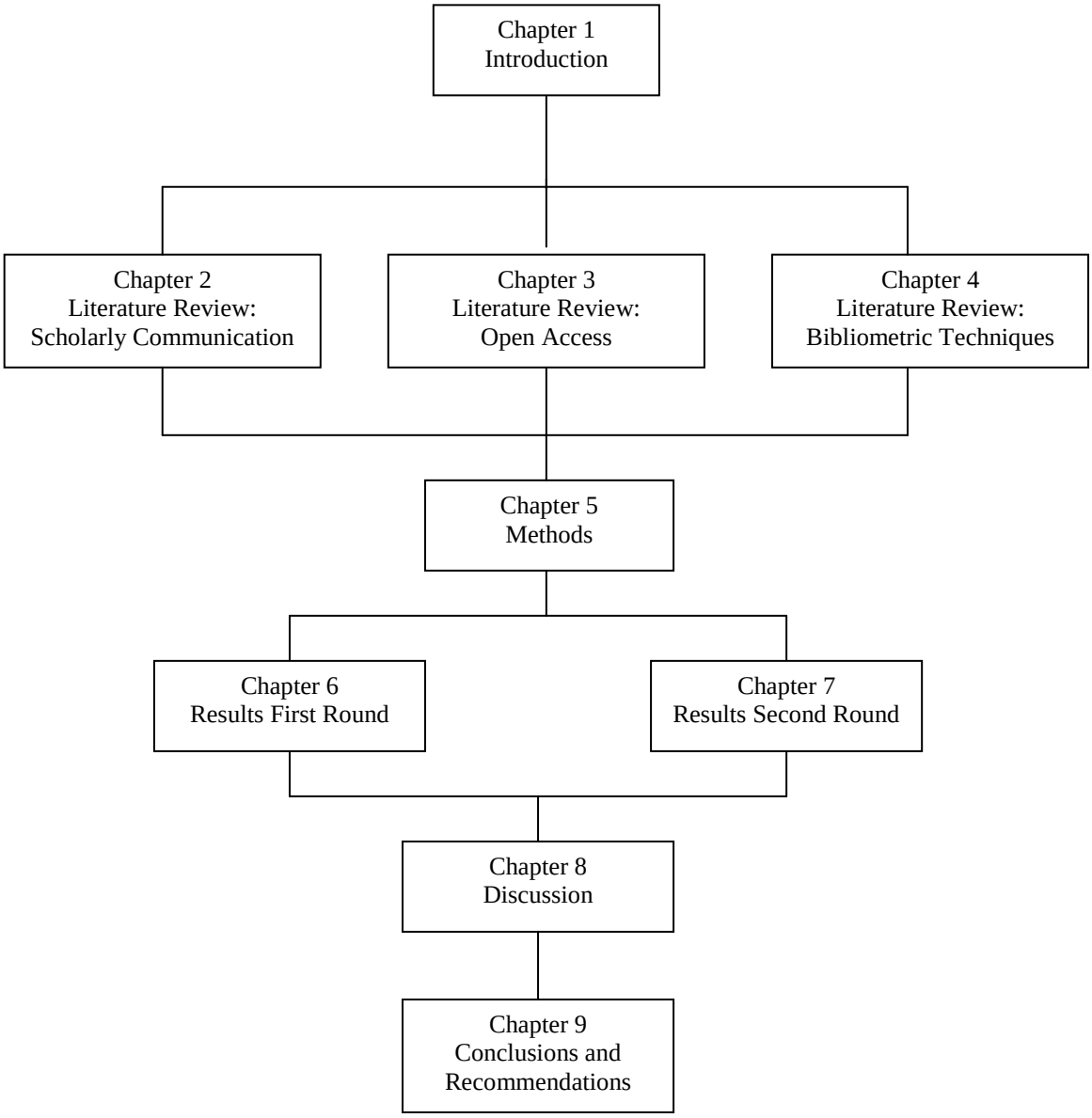
Objective 2: Confirm that the results found in Objective 1 were not a chance event.

Objective 3: Ascertain whether the OA/TA citation advantage is randomly evident in a population of journal articles and whether there is an early access advantage evident from OA articles in terms of patterns of earlier citations.

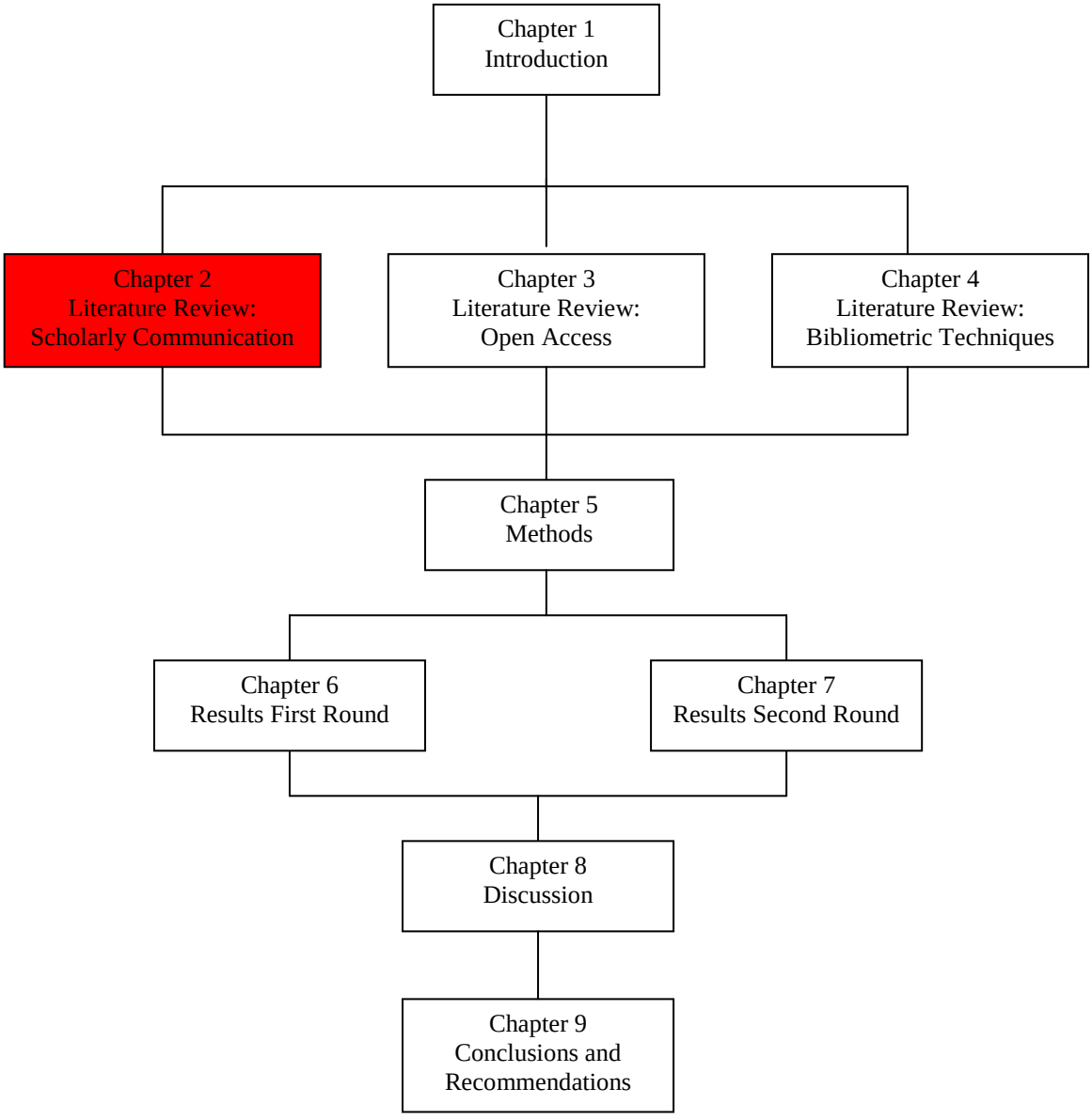
Objective 4: Determine if lower impact journals have a similar distribution of OA/TA articles and citation characteristics to their high impact counterparts.

Objective 5: To determine if causal links can be found between the OA/TA status of an article and the collective bibliographic details of the records associated with that status.

The following chart shows the structure of the thesis and how each chapter relates to the thesis as a whole.


Chapter 1
Introduction
Chapter 2
Literature Review:
Scholarly Communication
Chapter 3
Literature Review:
Open Access
Chapter 4
Literature Review:
Bibliometric Techniques
Chapter 5
Methods
Chapter 6
Results First Round
Chapter 7
Results Second Round
Chapter 8
Discussion
Chapter 9
Conclusions and
Recommendations

Chapter 2 Literature Review: Scholarly Communication



2.1. Introduction

Lamb (2004, p.144-145) considered that a definition for OA is still a work in progress. There are and have been many definitions, statements and declarations made about the OA movement to date (Suber, 2005a). Swan and Brown (2004a, p.4) do not attempt to define OA in terms of a mechanistic process, but describe it in an academic context, where the freedom of unfettered scholarship is paramount:

> OA – free access to scholarly information – underpins the core tenet of academic endeavour, which is the unfettered sharing of research communication. This core tenet permits the free exchange of ideas, results and discussion and encourages and accelerates scholarly achievement in every field.

The OA movement has grown out of the belief that the results of publicly funded, peer reviewed research, mainly in the form of journal articles, should be made freely available to those who would like to read them (Harnad *et al.* 2004, pp.310-314). International declarations, statements of support and numerous reports have agreed with this (see, e.g. Suber 2005a). At minimum, two main benefits are seen to accrue from OA: the potential research impact of academic articles is not lost, and anyone who wishes to do so can access them (House of Commons Science and Technology Committee 2004, pp.9-10). Willinsky (2006, pp.xii-xiii) finds two factors which he thinks have given the movement impetus and direction: the excessive and rising cost of journals and the publishing/archiving potential of the Internet and the World Wide Web. The OA movement has not, of course, developed in isolation; it sits, albeit somewhat uncomfortably at times, within the framework of an existing process of scholarly communication. The continuing good relationship between journal publisher and author is important and the two are, of course, mutually dependent upon each other (Rowland 2005a, p.3), the former for a means of reaching their audience and the latter for the content that makes publishing a profitable journal possible. The movement is also dependent for its success on many of the existing features of the current model of scholarly communication, for example editorial and peer review processes, and the willingness of authors to continue to give their work freely to journal publishers with few restrictions. Technological advances have also made OA possible; without them, the possibility of OA in its present form would be unlikely (Lamb 2004, p.146).

Some of the proponents of OA are at pains to make the point that it is concerned about gaining access to the results of publicly funded research, and that it does not, necessarily, prescribe any particular publishing model. (Harnad 2005c) Rather, it might be seen as a particular model of scholarly communication where OA provides another or secondary means of access to work that has been published, thereby hopefully increasing its impact. The difference is subtle but important and has been the subject of intense discussion (Guedon 2006). This said, the term OA publishing is frequently used to explain how particular publishing models might make OA possible (Schwartz 2005, pp.488-489).

Whilst the two concepts, OA and OA publishing, can be readily separated and understood, uncertainty about the impact of OA on the publishing industry fuels fears of loss of profit and market share with all the effects that this would have on the global industry and those employed in it (Morris 2005, pp.1-4). To fully appreciate the place that OA has in the model of scholarly communication and its potential impact, it is necessary to understand the elements of the current model of scholarly communication on which it would depend to be successful.

Given that until academic research has been successfully communicated it cannot be regarded as having any real value (Harnad *et al.* 2004 p.311), this review will open by looking briefly at the origins and growth of science and scholarly communication. It will continue by looking at the basic mechanics of scholarly communication, particularly peer review, models of publication and the importance of publishing to the individual and to research in general. The way in which journal publishing has been commercialised and the impact that this has had on library budgets will also be considered. The review will then move on to examine in more detail the structural components, developments and debate that have made it possible for an OA movement, potentially at least, to flourish. Finally, the review will look at the principal bibliometric tools that can be used to measure scholarly communication and the research which purports to demonstrate that indeed OA articles can potentially have more impact than their TA counterparts. Thus, the review considers OA within a broad context, showing how the movement has grown and the processes on which it is dependent, how it has influenced its own progress and considering the evidence that it does in fact increase research impact.

2.2. Scholarly communication

When those involved in research have their work published in academic journals, they are making a formal statement of their contribution to their discipline (Wellcome Trust 2003, p.1). They may have discussed their work amongst their peers, presented outlines of it at a conference, or posted preprints of it to a disciplinary or institutional archive, but, in the sciences particularly, it is the report of conference proceedings or more likely the journal article that will provide an authenticated enduring record of a researcher's work. For the author, the article formally establishes them as the originator of the work (Wellcome Trust 2003, p.1). The process of publication ensures, it is hoped, through peer review and editorial processes, elements of quality and integrity in the finished work. Finally, publication is of course the means by which the author's work is disseminated to their peers. Once the work has been published, fellow academics can discuss the work; they may agree or be critical of it, and they may advance the work by carrying out further research. Subsequently, other authors will publish their findings and the research will advance a stage further. This process is the basis of scholarly communication. More formally, the Association of Research Libraries (Association of Research Libraries 2000) describes scholarly communication as the:

> …formal and informal processes by which research and scholarship of academic staff, researchers and independent scholars are created, evaluated, edited, formatted, distributed, organised, made accessible, archived, used, and transformed.

2.3. The origins of the academic journal

Whilst nowadays the process of scholarly communication through the academic journal is regarded as routine to practising academics in most disciplines, the foundations of this form of scholarly communication saw its origins in the seventeenth century. Galard (2001, pp.3-5) suggests that René Descartes occupied an important place in the history of thought as he fostered a different approach to the study of the world and of science. The change in which Descartes was instrumental, Galard (2001, p.1) argues, was in bringing about the transition from the medieval scholastic tradition to one that was supplanted by the modern scientific mind. As Gribbin (2003, pp.107-117) shows, Descartes was not alone; with the astronomers Copernicus, Kepler and Brahe, for example, as his

antecedents and Galileo as a contemporary, the use of reason and evidence to explain the world had already begun. In this transition, the work of the nascent scientist was becoming increasingly based on observation and experiment, where the examination of evidence was the yardstick of scientific development, rather than conjecture and mysticism (Gribbin, 2003, pp.68-69). In this development, Descartes, like many of his contemporaries, became a member of 'invisible colleges', where like-minded scholars discussed their work through informal networks that were not part of any established institution (Tenopir & King 2000 p.56).

In the latter part of the seventeenth century, scientists were increasingly discussing and sharing, by communicating amongst themselves, their findings and their work. Of course, the invention of printing had made it possible for scientists to publish books, often anonymously, well in advance of the first journals (Meadows, 1998, p.3). However, scientists, aided by the development of postal systems, had, in the process of letter writing, a much more immediate and direct form of communication (Meadows, 1998, pp.3-5). In England, as in Europe, there was a significant growth in the number of people engaged in scientific research and in the formation of scientific organisations to foster such work.

In London, the Royal Society was formed in 1662 with just such an ambition and had as its secretary Henry Oldenburg, who began acting as a "clearing-house for information on new ideas and research" (Meadows 1998, pp.5-6). This involved Oldenburg in an attempt to keep up to date with new developments by engaging in extensive correspondence with other researchers. This process, quite apart from becoming too burdensome to manage, lacked formality and authority. In 1665, Oldenburg, aware of the French publication *Le Journal des Scavans*, suggested a similar journal for the Royal Society; hence *Philosophical Transactions of the Royal Society of London* came into existence in that same year. While the two publications were journals, only the latter, Guédon (2001) claims, was helping to "validate originality". In this process, Oldenburg and the Royal Society were setting in place the basic characteristics of the formal communication of science. Notably Oldenburg took the financial risk of publication himself, being effectively the first commercial journal publisher. Amongst these are the establishment of the priority of research amongst competing scientists, the process of peer review and the dissemination of a journal article as a durable public record of the author's work (Cox

2003, p.11). In this process, Guédon (2001) believes that Oldenburg and his Philosophical Transactions have cast a "long shadow" over the development of the academic journal, in that he effectively placed the academic journal at the heart of scientific communication.

2.4. The growing number of journals and scientists

Counting the number of journals which are current and being published is a difficult task; Henderson (2002, pp.133-136) considers the process is fraught with difficulties of definition and varying statistics. What is certain and undisputed is that there has been a steady rise in the number of journals since the seventeenth century. *Ulrich's Periodicals Directory* (2008) lists 26,850 peer reviewed journal titles available worldwide, with 16,587 of these available in an online format. Of these, 1825 are listed as OA, with 256 being indexed and listed in the *Web of Science* as core journals; all are available online. From these refereed journals, Harnad in 2004 (2004a, p.63) estimated that there was something in the order of two million articles written annually. This large number of journals needs many authors to provide their content. In *Little Science Big Science*, Price (1963, p.8) analysed the growth in science and the number of scientists associated with that growth. Working with data from the mid-seventeenth century onwards, he calculated that the number of scientists was growing at an exponential rate, with a doubling of their number every fifteen years. This led Price (1963, p.11), to the conclusion that there were "about seven scientists alive for every eight that have ever been". From this, it is evident that there are more active scientists working now than at any time in the past and hence unsurprisingly there are likely to be more articles being written in any one year than in any one year in the past. Given this, it is not surprising that there has been a commensurate growth in the number of journals in which scientists can publish their work. If Price's (1963, p.19) growth rate were to continue he speculates that:

> It is clear that we cannot go up two orders of magnitude as we have climbed the last five. If we did, we should have two scientists for every man, woman, child … in the population, and spend on them twice as much money as we had. Scientific doomsday is therefore less than a century away.

Price was, however, aware that such growth is unsustainable, or at least not beyond the middle of the twenty-first century; statistics from the USA seem to confirm that this is the

case. A snapshot of the growth of scientific occupations (life, physical and social science) in the USA shows the projected overall rise for the period 2004-2014 to be 16.4%, a significant slowing down in the increase when compared to the latter half of the twentieth century (Heckler 2005, pp.84-85). Computing and mathematical sciences have shown a stronger growth than the life, physical and social sciences. Taken separately, the growth rate of the life, physical and social sciences is higher than other sectors of the US economy, but "not by an order of magnitude" (Science and Engineering Indicators… 2004).

Whilst it cannot be disputed that the growth in the number of journals and scientists has increased dramatically, an exponential rate is perhaps an overstatement. Meadows (2000, pp.89-90) shows in his examination of the growth of scientific literature that there are also other factors which have augmented this growth. He points out, for example, that there has been a general increase in the number of articles published in any one journal and that this has naturally led to a commensurate growth in the number of pages in most journals. Whilst journal numbers have increased, in many cases their frequency of publication has also increased. Meadows therefore suggests that article counts are probably more meaningful than journal counts alone. He also argues that this increase in journals exists alongside a steady attrition in their number by an ever-present death rate amongst them, both in terms of their complete loss in the merger of journals and the waxing and waning of interest in particular fields of studies (2000, pp.91-100). Tenopir & King's (2000, p.242) research has shown that in America alone, the number of academic journals published rose from 4175 in 1975 to 6771 in 1995. This fact seems significant until it is noted that the number of scientists also doubled during this period, so the number of journals per scientist actually fell. Taking the statistics as a whole, however, they can, like Meadows, demonstrate an absolute growth in articles and pages published over the period (Tenopir & King 2000, p.242).

It is evident, then, that up until the end of the twentieth century at least, that more journals were being published than ever before, given that the birth rate of journals easily outpaces their death (Henderson 2002, p.133). Henderson (2002, pp.148-155) also discusses the effect of events in the twentieth century on the rise and diversity of journal publishing; major events, from the Second World War to space exploration and the Cold War, all increased competition between nations. With this competition spurring research

productivity and with generous funding from governments, the demand for ever more specialist journals increased, as did the volume of articles being written. This increase is amply demonstrated by the research of Mabe and Amin (2001, p.155) who noted that between 1900 and 1945 journal numbers increased annually by 3.3% but that this rose to 4.68% between the period 1945-1979, subsequently reverting back to almost their former level at 3.31%. Mabe and Amin (2001, p.160) also concluded from their overall figures that an "increase of about 100 refereed papers per year world-wide results in the launch of a new journal" which "suggests a growth in the potential annual author community of about 100-150 per annum for each new journal". In a slightly later analysis, Mabe (2003, p.196) is less certain of the continued growth in the number of academic journals, given that the growth of new journal titles "is starting to fail to keep up with the extrapolated 3.26% rate". There is recognition by Mabe (2003, p.196) that journal growth is related to the growth in the number of researchers, and given the projected slowdown in the number of researchers in the USA, this may be a limiting factor. It is also possible that this is simply a problem of the counting and definition of academic journals.

Whilst the number of journals published may have risen, their individual circulation figures as a whole have seen a steady decline over the last ten years. This evident decline results from a number of causes and is related to the particular nature of academic publishing and the circumstances in which it operates. The strongest correlation in this decline occurs between the cost of journals and the funds available to purchase them (Tenopir & King 2000, pp.33-35).

2.5. The growth of commercial journal publishers

As mentioned earlier, Oldenburgh published *Philosophical Transactions for the Royal Society of London* originally as a private commercial venture; it was only after 1752 that the society itself took financial responsibility for the journal (Henderson 2002, p.138). *Philosophical Transactions* was published thereafter to disseminate findings and not with the primary aim to produce profits; rather any surpluses would be used to help fund the purposes of the learned society. In this general model, income was derived from the society's members who, in return for their membership fee, received a copy of the journal. Non-member subscriptions from institutions and individuals added to the publishers' income. In the intervening years to the mid-twentieth century, learned

societies and university presses in the UK between them produced the majority of journals based on this not-for-profit model (Rowland 2005a, pp.5-6).

After the Second World War, as noted above, research funding expanded and commercial publishers could see opportunities to develop their interests in the academic market and increase their market share (Fredrickson 2001a, pp.64-65). Learned society publishers who were cautious in their approach to starting new journals helped, through their inaction, this commercial expansion. Fredrickson charts the rise of the Dutch company Elsevier as an international publisher towards its pre-eminent position today as the leading commercial publisher. As Fredrickson points out, Elsevier was:

> …moving away from publishing society-owned journals. Opportunities to start new journals were now offered by what was known as the 'twigging' effect in science, when sub-areas of major fields were receiving so much worldwide attention that the need arose for highly specialised journals in these subjects. As well as this, ventures into areas such as earth sciences, pharmacology and other biomedical areas took place on a larger and, from 1960 onwards, planned scale. (2001a, p.67)

Elsevier found, together with other publishers like Pergamon and Blackwell, that they "really did not need to market their publications" (Fredrickson 2001a, p.69); such was the demand, they effectively sold themselves. Understandably, learned societies did not necessarily see their role as potential commercial publishers where they might have used their surpluses from journal subscriptions to launch further profitable journals in their particular field (Wellcome Trust 2003, p.16). Rather like the Royal Society, they wanted to use their surplus funds to advance the aims of their society (Cameron 2001, pp.247-248). Sarkowski (2001, pp.25-33) suggested that there was also a similar pattern of publication elsewhere, apart from in Germany, where the commercial publishers had, until 1939-1945, dominated the market for academic journals and books.

2.6. Commercialisation of journal publishing

As the journal publishing industry's structure changed during the second half of the twentieth century, large commercial publishers were collectively coming to dominate the market where formerly the not-for-profit publishers had done so. Commercial publishers

took the opportunity to fill the gaps left by society publishers and the struggling post-war German publishers, particularly where there was a demand for more specialist journals which the learned societies had declined to meet (Cameron 2001, pp.246-252). As their market share increased, the commercial publishers were gradually able to charge a market rate for their products, unhampered by the more conservative aims and objectives of learned societies (Cameron 2001, pp.248-253). Through this process, commercial publishers, collectively, have come to dominate the academic journal market. In so doing, they have often been accused of making undue profits, but Atkinson in an interview with Duranceau, makes a realistic, if somewhat reluctant, acknowledgement of their role and of the conflict this gives rise to:

> A commercial publishing operation is first and foremost a business. If it is a public company, its absolute primary goal must be to achieve the very best possible return on investment for its shareholders. The primary goal of scholarship and research is to advance knowledge and to ensure that such knowledge is available to all who are in need of it… These are very different goals. (Duranceau 2004, p.128)

The commercial position has been strengthened by consolidation amongst these publishers some of which have comparatively recently either merged or been taken over by larger companies, thus concentrating more and more of the market share in fewer hands (Wellcome Trust 2003, pp.20-21). Reed Elsevier, the largest supplier of scholarly journals, for example, publishes over 2600 such journals (Reed Elsevier…. 2008).

While these significant structural changes were taking place, the scholarly academic journal was still being produced and sold effectively as it had been since its origins in the seventeenth century, on paper and using the same basic publishing model (Peek 1996, pp.4-7). In the latter part of the twentieth century, however, technology made it possible, through the development of the electronic journal, for publishers to devise new models of access and delivery of their journals to institutions and readers. This change has had a fundamental effect on the way publishers can market and sell their journals. Similarly, this new technology, with the possibility of separating journal content from its paper medium as well as the possibility of remote access, also opened the way to the OA movement (Willinsky 2006, pp.xi-xiii).

2.7. Journal costs

The cost of a printed journal to a subscriber depends largely on the number of copies that are sold. This seems to be a self-evident fact, in which the greater the potential circulation, the more the costs of sales and production can be spread across the subscription base. First copy journal costs to the publisher are however, very high; that is, the cost of producing the journal initially as effectively a single issue which carries all the fixed costs associated with editorial processes and the cost of sales (Tenopir & King 2000, p.247). First copy costs for an electronic journal are similarly high; however, the difference in cost of distributing to 100 or 1000 subscribers with an online electronic journal is much lower than the incremental cost of printing and distributing a printed journal. This said, the capital investment in technology to deliver, archive and manage journals electronically has to be made in the first instance, along with supporting the technical expertise required to maintain it (Houghton 2005, pp.171-172). The journal market has some particular characteristics which affect access and cost (Houghton 2005, pp.165-169). Unlike an item or object that can be bought and consumed only once, the contents of a journal may be read and subsequently read by another reader; the information the journal contains remains in its original form and is not consumed in an industrial sense. Houghton further considers that this means:

> …that ideas and information exhibit very different economic characteristics from the goods and services of the industrial economy, and that the social value of ideas and information increases to the degree they can be shared with and used by others. The more such information goods are consumed, the greater the social return on investment in them. (2005, p.169)

Information in this case has to be read before its value to the consumer can be measured. Given the potential for uncertainty about the quality of an article in a journal, the reader generally looks for clues that will help them decide whether they wish to purchase it. Readers seeking clues as to the worth of a journal will look to find measures of quality that relate to the standing of the journal in terms of its editors, its contributors and perceptions of its readership and image, or even its impact factor. These factors, Houghton (2005, p.169) concludes, are an important determinant of the price a publisher may charge for a journal.

Many journals occupy a unique position in the discipline in which they publish. As new research areas emerge, new societies have been formed to address the needs of that particular discipline. Meadows (1998, p.21) cites the example of computer science, where societies have been formed and specialist journals have been published, either by the society itself or by commercial publishers, to meet the needs of society members. Often, journals are unique to a particular discipline, making it very difficult for a researcher to work in that field without having access to them. Additionally, hierarchies are established where there is a perception by the readership, and often by employing institutions, that certain journals lend more credibility to published research than others. Such perceptions and the uniqueness of some journals in their field, have allowed publishers to create a monopoly where the journal cannot easily be substituted by another of equal standing or content (Houghton, 2005, pp.169-171). In these situations, the cost of a journal, depending on who is publishing it, may be more a matter of its value to its readers rather than reflecting the actual production cost, and thus publishers may be in a position to charge whatever they believe the market can stand. Rowland (2005a, p.6) suggests that these higher costs are a function of the type of journals that commercial publishers produce, that is, highly specialised, low circulation titles and of course the need to generate profits.

There are a large number of academic journal publishers, the majority of which are small not-for-profit society publishers, producing a handful of journals each. The large commercial publishers, on the other hand, collectively publish the majority of academic journals, and, being few in number, effectively control the journal market in terms of prices charged. In a report by Cox (2003 p.21) for the Association of Learned and Professional Society Publishers (ALPSP) which surveyed a cross section of publishers, both commercial and not-for-profit, it is clear that the larger the publisher the greater the number of new journals they will publish. In all respects it appears that costs at commercial publishers are greater than those found at society publishers. Thompson (2005, p.101) reports a study which compared the cost of the top ten most cited commercial and not-for-profit economics journals and found that the page costs for the commercial journals were six times higher than the not-for-profit journals. Similarly Page, Campbell and Meadows (1997, pp.6-7) also show that commercial publisher costs are significantly higher and that the rate of increase of these costs is also higher.

Additionally, they showed that the number of subscribers to new journals launched by commercial publishers has been in steady decline; where a new journal may have achieved a thousand subscribers after five or six years of publication, this number had declined by half during the period 1988-1993.

2.8. The serials crisis

Given the growth in the number of journals being published and the concentration of journal publishing in the hands of large commercial publishers, and the increase in serial prices, it is not surprising that King and Tenopir, (2000, pp.253-254), identified the rising cost of journals as the most important trend (financially at least) to occur in journal publishing in the previous forty years. The last twenty years, they argued, initiated the following events that led to spiralling journal costs and hence the now well recognised 'serials crisis':

- Personal subscriptions [to scholarly journals] have dropped to half the level of twenty years ago
- This drop in personal subscriptions prompted publishers, in an attempt to maintain profits, to increase their charges to libraries at a rate far higher than inflation.
- With reduced budgets, libraries have cancelled subscriptions to non-core journals and, mainly at the expense of book purchasing, have tried to maintain their coverage as best they can.

A cycle of lower-than-inflation budget provision for libraries and these increasing journal costs has led to further journal cancellations, which in turn has reduced the subscription base and income for publishers. If publishers wish to maintain their current profit margin, they must in turn raise the cost of the subscription to their existing customers, and thus the cycle is perpetuated. Guedon (2001) firmly places the blame for this 'serials crisis' on the shoulders of large commercial publishers who, he maintains, have ruthlessly exploited their monopolistic position. He also believes, however, that "scientists simply need these publishers too much and they cannot conceive of an alternative way to manage a successful career" (Guedon 2003, p.139). The journal market is often described as being inelastic: the Wellcome (2003, p.15) report on publishing describes "price-elasticity as low", in the following words:

> …readers will not normally be much influenced by price in their decision whether or not to read an article. Demand is relatively unresponsive to price. A primary reason for this is that journals are not close substitutes for each other. [hence] A specialised journal thus acquires a significant amount of monopoly power. Readers are unable to find alternative sources.

Guedon (2003, p.139) sees the academic exacerbating this problem in two ways: as readers, they insist on having access to journals irrespective of their cost, and as authors, they want to publish in the most prestigious journal possible, whatever its cost. Not only this scholars find themselves under pressure to publish more to justify their tenure and meet the requirements in the UK for example, of the Research Assessment Exercise. As Steele (2005, p134) notes, for the large publisher at least there is no serials crisis as such; profits remain high, with operating profits in the range of 33-35%, but the crisis resides firmly with the institutions who have to buy them.

Interestingly, despite the ever-increasing cost of journals, the unit cost of holding and managing electronic journals for a library is significantly cheaper, compared to their paper counterpart (Montgomery & King 2002). The first copy costs (Tenopir and King 2000, p.39) remain just as high in whichever format the journal is destined to appear. However, giving wholesale access to electronic journals is cheaper than supplying them in paper form (House of Commons Science and Technology Committee 2004, p.41). These two factors made it possible for publishers to develop online business delivery models that allowed for the 'bulk supply' of journals to libraries. To enable this, publishers developed electronic site licences which allowed them to manage access to their journal collections in discrete lots on an institution by institution basis (Thompson 2005, p.100). This allowed publishers to aggregate their journals into marketable lots, which they could then sell to institutions. Two models of supply were developed from this process, the first called 'bundling' and the second the so-called 'Big Deals'; both allowed in different forms the bulk sale of journals to institutions. These two models are discussed in more detail in section 2.9 below.

The effect of these two new supply models was to reverse the upward trend in average prices paid for serials; the trend is now downward and has been so for the last few years (Library and Information Statistics Unit 2005, p.4). This fall in average prices in the UK has been directly attributed to the 'bundling' by publishers of large numbers of journal

titles together. These bundles are offered to libraries or to a consortium of libraries even though they may contain journal titles to which the libraries do not wish to subscribe. The cost of a bundle is often substantially cheaper than the individual price of each journal if subscribed to separately, hence the average cost of the journals falls. The data also shows, ironically, that despite the fall in the average cost of serials, increasingly more of the academic library acquisitions budget is being spent on serials and electronic resources (Library and Information Statistics Unit 2006, p.139). This section of the budget increased as a proportion of the total spends by 11.3 per cent and 33 per cent respectively in the five years up to 2005. The serials budget for the whole of the UK HE sector for 2005 was consuming 52.4 per cent of the entire library acquisitions budget. The main casualty in this funding change has been in the purchase of books, whose share of the budget declined by 19.3 per cent over the same period. The Association of Research Libraries (Kyrillidou & Young 2008, pp.12-16) report similar findings for North America, where statistics for 2005-06 show that serials costs have been increasing much faster than inflation with serial expenditure having risen by 321% over the last twenty years, compared to only 82% for monographs from a base line of 1986.

In response to these ever-increasing costs ARL, created an alliance, both in North America and Europe, of universities, research libraries and interested organisations. This is called the Scholarly Publishing and Academic Resources Coalition (SPARC). SPARC's agenda "focuses on enhancing broad and cost effective access to peer reviewed scholarship" (SPARC 2005). It achieves this through three routes: advocacy, education and helping to create competitive alternatives to the established commercial STM press. Notably, OA journals have been developed and promoted; *Organic Letters* appeared in 1999 and has become a successful direct competitor to the expensive flagship journal *Tetrahedron Letters* published by Elsevier. At $4332 for a European institutional subscription, *Organic Letters* is about half the cost of the former (ACS Publications 2008). Ironically, through its success *Organic Letters* has become an important journal for libraries to take along with *Tetrahedron Letters*, so instead of libraries believing they could purchase an alternative, cheaper journal, they now find they should be taking both (Bachrach 2001, p.137). This has inevitably put further pressure on library budgets.

2.9. Bundling and big deals

With subscriptions to journals falling and the consequent price rises, commercial publishers introduced, as noted earlier, the concept of 'bundling'. In this model of journal delivery, publishers supply journals in an arrangement "whereby print and digital formats are provided as a bundle, often with all digital journals bundled together as a single product" (Wellcome Trust 2003, p.5). This general strategy has been analysed by Bakos and Brynjolfsson (1999, pp.1613-1614), who find that it is more likely that "a multiproduct monopolist will extract substantially higher profits by offering one or more bundles of information goods than by offering the same goods separately". This follows, they argue, from the fact that there will be less variation in the valuation of a bundle of goods than would be the valuation of individual items from that same bundle by people (or institutions) who buy them (1999, pp.1625-1626). They go on to note that this strategy is particularly effective where the cost of delivery is effectively almost zero, as in the delivery of digital products. This has allowed publishers, Houghton (2005, pp.173-174) suggests, to sell their bundles at an average price through site licences, with these licences controlling access and delivery to the specified groups. In this way, Houghton continues, publishers have retained control as digitisation and online distribution have changed the economics of journal delivery.

Partly in response to these bundling deals, many institutions formed themselves into consortia as a way to give themselves a better negotiating position with publishers (Guedon 2001). Whilst this has given institutions more collective power when negotiating with publishers, publishers have maintained their profitability by managing their pricing structures to achieve their business targets (Jeon-Slaughter, Hertkovic & Keller 2005). Despite this, as José and Pacios (2005, p.189) note, many libraries, particularly smaller ones, have benefited by reducing duplication and gaining access to much larger collections of journal titles than would have been possible if they had negotiated separately. Guedon (2003, pp.129-130), however, sees mixed results from these consortia, with limitations on the ownership, access and preservation of the journal titles in the bundles. The benefit for readers, however, who are probably either impervious to or oblivious of these issues, is an electronic service which delivers access to journals from their desktop (Cox 2003, p.14). Publishers offering 'Big Deal' access to journal titles have increased this accessibility.

The writers of the Wellcome report (2003, p.9) consider that the 'Big Deals' offered by publishers are different from the mixed bundles of paper and electronic journals described above. In a 'Big Deal', an institution may be offered, separately from any consortium agreement, large-scale access to a bundle of electronic journals for a fixed price and period. These bundles can be made up from a publisher's entire collection or a particular part of a collection which covers a discrete discipline – a subject bundle (Cox 2003, p.6). Such deals usually tie the institution into a fixed term contract where costs are fixed with little or no room to cancel individual journals within the package (Gibbs 2005, p.91). Frazier (2001) describes "the Big Deal [as] an online aggregation of journals that publishers offer as a one price, one size fits all package" The effect of this large scale access, Frazier continues, is that "It bundles the strongest with the weakest publisher titles, the essential with the non-essential". In her experience, Gibbs (2005, p.91) as a librarian, found that while some of her readers did access some of the 'new' journal titles that the Big Deal had given, their original core journal holdings remained the most important and the most accessed. Gibbs (2005, p.92-93) and her colleagues found that it was difficult to cancel and substitute new journals within the Big Deal as research interests changed, and that this led eventually to cancellation of contracts with Elsevier and Blackwell.

An analysis by the publisher Emerald, however, of their 'Management Xtra' package found that usage was more dispersed than expected. Countering criticisms that there is too much redundancy of usage in Big Deals, Evans and Peters (2005, pp.155-156) showed that the Pareto rule of an 80/20 usage profile did not apply. They calculated from 2004 download statistics that 80% of the usage came from 47.4% of the titles in their bundle. Evans & Peters compared their results with a consortium of university libraries in Spain and found that the Spanish download statistics were very similar to their own, making their findings closely comparable (2005, pp.156-157). Not dissimilar findings were made by Nicholas *et al.* (2005, pp.252-253), who analysed, by deep log analysis, journal usage from the OhioLINK consortium, the first and original Big Deal. They found that "The idea that Big Deals offload unwanted journals on libraries (and users) certainly seems to have been disproved". From the total of 5872 journals available on OhioLINK, 5193 (88%) had been accessed to view an article. Making a comparison with the work of Evans and Peters above, Nicholas *et al.* noted that "half of all journals accounted for

about 93% of usage". However it should be also noted that just 10 per cent of journal titles accounted for 53% of the usage. (Nicholas *et al.* 2005, p.253)

2.10. Online access and measuring usage

An essential component of the OA movement is the availability of electronic online access to the research article (Willinsky 2006, pp.xi-xiii). Electronic journals in significant numbers have been available since the-mid 1990s (Rowland 2005a, p.7). In the survey for ALPSP, Cox (2003, p.23) found that three quarters of journals published were available online. Of those journals listed in the *Journal Citation Reports*, 94% can be found online (Ulrich's Periodicals Directory 2008). This is an upward trend and is expected to continue: the report for the Wellcome Trust (2003, p.22) found that academics, librarians and publishers:

> …were unanimous in seeing electronic access as the major source of articles for the research community in the future. Paper journals were seen as playing an important though probably, diminishing part. Academics have become used to the convenience of electronic access at their desktops.

Similarly Willinsky (2006, p.15) found in a survey that the reading culture and expectation of students and university staff "ranked online journals ahead of books, print journals and other resources", with the consequent effect that his library expects eventually to move to an online environment only for their journals.

Judging the quality, saleability and usage of a journal is difficult to determine. Page, Campbell & Meadows (1997, pp.136-138) give a number of strategies as to how this might be achieved. The counting of citations to journals and articles is one measure of their usage and decay, but this only records the usage of a small number of users, in this case authors. Online access to electronic journals has, on the other hand, progressively allowed something which formerly was very difficult to do, that is, to measure with any certainty the usage of journals by their readers and how they are accessed (Huntington, Nicholas & Watkinson 2004, pp.249-250). They believe that access to, and analysis of, the user logs, which record the computer transactions of readers while they search for relevant articles, is very important because:

> …they are a direct and immediately available record of what people have done: not what they say they might, or would do: not what they were prompted to say: not what they thought they did. The data are unfiltered and speak for themselves…(2004 p.251)

What is clear is that the ability to count and understand usage by institutions and publishers is almost certain to affect the way journal packages are sold to libraries and consortia (Nicholas *et al.* 2005, p.256-257). This counting has certainly informed the way some institutions now view the journal products that have been sold to them, resulting in the cancellation of some of the commercially bundled packages they have bought (Duranceau 2004, p.128).

2.11. Peer review

Peer review is critical to the quality and authority of the academic journal (Harnad 2004a, pp.64-66). To be credible within a research community and eventually to the public, accessible articles need to have reached a perceived standard of quality. The process of peer review and editorial control can give both the author and journal authority and status. De Vries (2001, p.231) opens his analysis of peer review by summing up the importance of peer review to the success of science and those who participate in it by suggesting that:

> Success in science depends as much on the scientist's ingenuity as on recognition of his or her achievements by colleagues. Positive evaluation of established results by one's superiors is the basis of any career. Published reports of recent findings, clearly having survived critical examination by outsiders, are its stepping stones. The sum total of admitted publications is the canonization of science. Peer review is the name of the infrastructure which supports the system.

For Weller (2001, p.16), a working definition of a peer-reviewed journal is:

> …one that has a portion of submitted manuscripts evaluated by someone other than the editor of the journal.

and its purpose is "to protect the integrity of science and scholarly communication" (Weller 2001, p.321).

For Henderson (2002, p.156), the "integrity of the scientific record is the burden of science editors aided by peer review". Peer review, as noted earlier, began in the seventeenth century, when the *Philosophical Transactions of the Royal Society of London* started. It was suggested by its board that before any article could be published by them, it should first be reviewed (Cox 2003, p.11). From these beginnings, the process has evolved into a recognised formal requirement and is now accepted as the routine procedure for those who wish to publish in a scholarly journal (Weller 2001, p.3).

The purpose of peer review, as seen, for example, by the *Journal of American Medical Association*, includes ensuring that the content of an article is original, the data are valid, and that the conclusions are reasonable and justifiable (de Vries 2001, p. 235). Harnad (1996, pp.109-111) describes the process as one of quality control, where the quality of the article is assured by a rigorous process of checking and verification by those qualified, and recognised as being qualified and authoritative in their field of expertise, to make such judgements. Once undertaken, the process should ensure, for the reader as the consumer of the work, an accurate and proper representation of the research undertaken and that the conclusions drawn are an accurate reflection of the work carried out. Secondly, journal publishers are looking to ensure that the standards set by the editorial board are being met, so that the perceived quality and standing of the journal is maintained in relation to other journals in the same field. Those attempting to publish, for example, in *Nature* or *Science* will find that rejection rates are high, (Meadows 1998, p.183), as is the perceived reputation of both journals to their authors and readers. Although Weller (2001, p.71) has not found conclusive evidence that there is a relationship between rejection rates and the importance of a journal, nor that those journals who reject the most articles publish the most important articles, the perception is that the two go hand in hand. Weller (2001, pp.17-24) lists the differing rejection rates of many core journals; they range between 11 and 95 per cent and whilst the perception is that the rejection rates are higher in the humanities, she found no trends that would consistently support this conjecture. By contrast, Meadows (1998, p.184) confidently asserts that rejection rates amongst leading journals are "clearly lowest in the sciences and highest in the humanities, with social sciences in between".

In the process of peer review, authors have their manuscripts reviewed, usually by external peers, to establish whether their work meets the requisite academic standards

demanded by the journal in which they wish to publish. An article may be accepted, accepted with revisions, or rejected; the editor may adjudicate if those carrying out the review disagree (Meadows 1998, pp.181-183). The process is important to the author and to the journal. For the successful author and the journal, reputations may be established or at the very least be maintained (Thompson 2005, pp.45-46.)

The process of peer review is, however, contentious and is often sometimes viewed with suspicion. It is blamed as being the source of delays, mismanagement and disputes about its efficacy. De Vries (2001, pp.239-240) reports the British Medical Journal's editors as stating:

> …that peer review is erroneously credited as being "a good method of keeping poor quality work from publication, whereas the evidence suggests that with persistence even the most flawed work will eventually find a home. Considerably less than five percent of papers in current journals contain a message that is both scientifically sound and relevant to doctors"

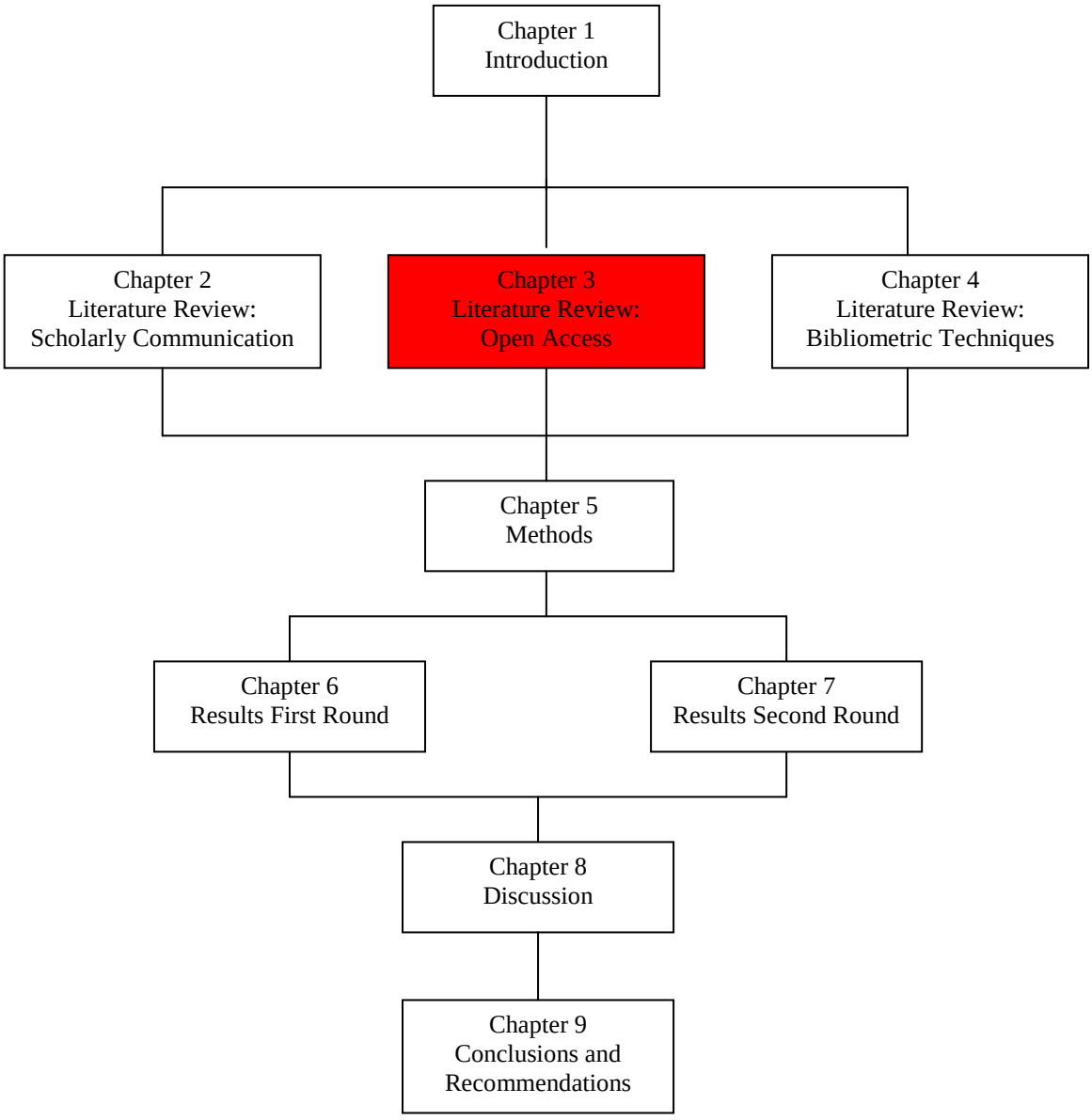
Meadows (1998, p.191) also reports on two studies which showed that with persistence, authors can get their work published. In a study that looked at articles rejected by the *British Medical Journal*, three quarters of those rejected were published elsewhere. The second study found that across a range of disciplines 60% of authors who had had their work rejected by their journal of choice were published elsewhere. In a similar, but earlier study Cronin and McKenzie (1992, pp.27-29) found that of 101 manuscripts which were rejected by the prestigious *Journal of Documentation* 28% of them were published elsewhere. The majority of which 'traded down' in terms of being published in a lower impact journal.

Ginsparg also finds the system somewhat flawed. Outsiders, he thinks, would soon:

> …learn that peer-reviewed journals do not certify correctness of research results. Their somewhat weaker evaluation is that an article is a) not obviously wrong or incomplete, and b) is potentially of interest to readers in the field. The peer review process is also not designed to detect fraud, or plagiarism, nor a number of associated problems…(Ginsparg 2003)

Weller (2001, p.322), however, sums up her analysis with the conclusion that "…editorial peer review is messy and does not always work as it should, but it is essential to the integrity of scientific and scholarly communication". Correspondingly, Davies and Greenwood (2003, pp.160-161) found in a JISC survey of specialist opinion that those involved in peer review, while agreeing that it was expensive and time consuming, believed that it remained the only sustainable model. Rowland (2002, pp.249-250), reviewing an ALPSP survey, also found that there was a fairly high rate of agreement between the authors and readers of electronic journals that peer review was important to them (81% and 80% respectively).

Chapter 3 Literature Review: Open Access



3.1. OA Movement – definition

OA and the objectives of the OA movement have been defined and stated many times, chiefly in manifestos, declarations and statements from various organisations in support of the OA movement. Peter Suber (2006), in his *OA News*, gives a straightforward account of the OA movement, describing it as:

> Putting peer-reviewed scientific and scholarly literature on the internet. Making it available free of charge and free of most copyright and licensing restrictions. Removing the barriers to serious research.

There are, however, a number of key definitions amongst the many, three of which Suber (2005a) contends have been the most influential on institutions, funding bodies and governments alike. They are the Budapest, Bethesda and Berlin statements. The first, taken in part from the Open Society Institute's Budapest OA Initiative (2002), states that:

> …By 'OA' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited…

The other two statements are very similar to the above. The Berlin (Berlin Declaration on OA…[n.d]) statement slightly modifies the wording used in the Bethesda (Bethesda Statement on OA Publishing…[n.d]) statement, which in turn was a refinement of the original Budapest statement. Both the Berlin and the Bethesda statements do however, share an emphasis, in that they both state that the author gives irrevocable access, and consents in advance to letting users:

> copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship…(Suber 2005a)

In practical terms, this means for the recently launched Public Library of Science journal *PLoS Pathogens* that its "quality content is available to read, copy, distribute and use without limits and at no cost" (Public Library of Science 2006) and "so long as the original authors and source are credited", (Public Library of Science FAQ…[n.d.]). At no cost, that is, to the reader, the cost of publication being supported by mixed funding coming from donations made directly to PLoS and payments from authors or authors' institutions for each article submitted for publication (Public Library of Science FAQ…[n.d.]).

3.2. Drivers to OA

Friend (2006), Swan and Brown (2004a, pp.4-6), Willinsky (2006, pp.xii-xiii) and Lamb (2004, p.146) identify between them a number of key drivers, which, they believe, have created a favourable social and technical climate for the OA model of scholarly communication to become a realisable possibility:

- the escalating cost of serials, shrinking library budgets and the impact this has had on limiting access to these serials for researchers, academic scholars and institutions;
- the development of the Internet, the World Wide Web and ubiquitous access to it;
- the use of the Internet in fostering new forms of storage, delivery and the sharing of electronic information products;
- the continuing growth in the output of science and the growing awareness of the users of this resource and those funding it, of the financial barriers erected around the dissemination of research findings.

On his timeline of the OA movement, Suber (2005b) has recorded many of the important steps in the progress towards OA. Amongst the many critical entries on this timeline, of particular importance is an entry from 1994, when Harnad posted his 'Subversive Proposal' to a discussion list hosted by the Virginia Polytechnic Institute. The proposal suggested, effectively, that the scholarly author had an alternative publishing option to making the "Faustian bargain of allowing a price tag to be erected as a barrier between their work and its (tiny) intended readership…" (Harnad 1994). That is, that the scholarly author could now make their work available to their peers by archiving it electronically to a publicly accessible archive, rather than simply publishing it through the traditional

academic press where access is controlled by payment. Harnad gave some explanation of the mechanics of this process, but most importantly, the concept and basis of an OA movement in a currently recognisable form had been proposed.

It was the development of the Internet and the World Wide Web which made this proposal possible, and without such an electronic infrastructure the posting of preprint or e-print articles or even the electronic journal (Tenopir & King 2000, pp.38-39) would not have been possible. Given that it is access to, and the dissemination of, the contents of the electronic journal, in whatever form, which is central to the OA model, Lamb (2004, p.144) asserts that "It is in the context of the full promise of electronic journals that OA has emerged as a new publishing model".

3.3. Pre and postprints

In the process of self-archiving authors are placing an electronic version of their work on either an individual or a departmental website or to a disciplinary or institutional archive where the work becomes available to anyone with Internet access. When authors do self-archive their work they may either archive a preprint or a postprint of their work, or both. Harnad (2003) explains that "Preprints are drafts of a research paper before peer review and postprints are drafts of a research paper after peer review". Not all preprints will be published. Postprints are in fact more frequently deposited than preprints, except in long-established archives such as arXiv (Swan 2005). The majority of publishers now allow authors to self-archive their postprints to an institutional archive, to a personal web page or a departmental web site (Sherpa Romeo 2008). Journal publishers may allow immediate self-archiving of a postprint or specify an embargo period after which it may be archived, and/or apply other particular conditions.

3.4. Models of OA

Harnad *et al.* (2004, pp.310-314) describe two distinct models of OA. The first is the 'green' route, where authors self-archive their work to an accessible repository, and the second is the gold route, where authors have their work published in OA journals. There can be shades of green and gold where the model adopted by authors or publishers is not pure, in the sense that it has been modified from its original into some other agreed form. Work by Gadd, Oppenheim & Probets through the RoMEO studies (Project

RoMEO…[n.d.]) encouraged publisher approval for authors to self-archive their work, subject to various embargo periods. The results are now hosted within the SHERPA/RoMEO (SHERPA…[n.d.]) web site, which gives details of the restrictions publishers have applied and the 'colour coding' which indicate these restrictions; of the 127 publishers listed, 93% now allow self-archiving in some form, either with or without restrictions. This type of arrangement is exemplified by the way authors are constrained in some cases only to self-archive their postprint versions after an embargo period. Oxford Journals' policy, for example, is that articles published in its journal *Brain* can only be self-archived after a twelve month embargo period rather than immediately after publication (Oxford Journals Access…2006).

A distinct version of OA is provided to developing countries, where publishers have provided access to their journals, often freely or at minimal cost, under some form of international sponsorship; two examples described by Rowland (2005b, pp.104-106) are the Health InterNetwork Access to Research Initiative (HINARI 2006) and Access to Global Online Research in Agriculture (AGORA 2005).

3.5. The green route

The green route requires authors to self-archive their peer reviewed postprints into institutional or subject based repositories as soon as possible after publication (Crow 2002). This route is favoured by Harnad, who sees it as the quickest way to get peer reviewed research online and available to all. This is an interesting transition for Harnad; the journal *Psycholoqy* which he started is arguably the first peer reviewed journal to be made freely available to all on the Internet and is an exemplar of the gold route to OA (Harnad [n.d.]). Harnad has subsequently observed that publishers see the gold route as risky and that self-archiving is more acceptable to them (Harnad *et al.* 2004, pp.313-314). Added to this, self-archiving to an institutional archive "enhances the visibility of the researcher's institution" and while he thinks the "golden route may prove to be the future for journal publishing, the road to maximising journal article access, usage and impact right now is the green road of OA self-archiving" (Harnad 2005c). The green route also allows for universal access to the results of publicly funded research, and appears to the reader effectively to be free, even though there are costs associated with the setting up and running of these archives. It is constantly argued, now that most publishers in effect

have given the green light for authors to self-archiving their work, that authors should take advantage of this and self archive all their postprint articles to institutional repositories (Harnad *et al.* 2004, p.313). Although this is in principle an effective strategy, not every institution has a repository into which academics can archive their work and authors in any case seen reluctant to make the effort. The issues around self-archiving are discussed in section 3.13

3.6. The gold route

Authors may publish their articles in OA journals that are free to end-users. These 'gold' journals have their costs met by means other than by subscription. Unlike the green self-archiving route to OA, in which authors may archive their work directly to a particular archive, the options for the gold route are more diverse. The Directory of OA Journals (DOAJ) defines an OA journal as one that "use(s) a funding model that does not charge readers or their institutions for access" (DOAJ 2008a). This directory, created and maintained by Lund University, provides a guide and access to such journals. The list currently holds over 3257 journal titles which are quality controlled scientific and scholarly electronic journals that have a conventional editorial or peer review process (Directory of OA Journals 2008b); this represents about 12% of the 26850 journals listed by Ulrich's as peer-reviewed journals. Users can access journal titles from the list electronically from wherever the journal is hosted.

Morris (2006, pp.73-76) and her colleagues analysed the content of the directory in early 2005 and found a number of journal titles which were defunct, not original or inaccessible; this represents 14% of the directory's content. It is fair to say, however, that DOAJ pursues an active de-listing policy, having removed 50 titles in the latter part of 2005 where editorial, publishing or access standards had not been maintained (DOAJ 2006b). It is significant to note that the peak in the growth rate of OA journals was at its highest in 2001 (from those listed at DOAJ) and has been in steady decline ever since (Morris 2006, pp.73).

Publishing in fully OA journals has been viewed with some suspicion by prospective authors. Swan & Brown (2004a pp.27-29) found in their survey that many authors were unfamiliar with OA journals and were not confident about publishing in them. The authors also perceived these journals to be low in impact, prestige and readership. This

impression was not helped by the report by Kaufman Willis (Association of Learned and Professional Society Publishers 2005a, p.49) for ALPSP, which suggested that OA journals were less rigorously reviewed than their subscription counterparts, that over 40% of them were running at a loss and that this situation was unlikely to improve. The report was criticised for these assertions (Baynes 2005; Friend 2005), and a subsequent post publication addendum was published, correcting some of these impressions. This was followed, however, by a complex argument, within the addendum on various definitions of peer-review, of models of its rigour and whether the quality of peer review was associated with the journal's business model. The conclusion was that more research was required (Association of Learned and Professional Society Publishers 2005b, pp.1-5).

Taking a more detailed look at funding models for OA, the survey conducted by Waltham (2006, p.16) examined "if and how learned publishers can consider making a transition to a sustainable open access model, and what the funding sources and requirements would need to be in order to do so". Looking at nine not-for-profit society publishers and a selection of the journals they published, Waltham (2006, p.2) found that bundled print and online subscriptions accounted for 29% of the circulation but 75% of the revenue in 2004, with print subscriptions falling by 43%. While these substantial changes in the subscription base were occurring, the number of articles submitted and published rose respectively by 35% and 25%, with an increase in pages published of 33% over the period 2002-2004 (Waltham 2006, p.3). These increases, Waltham (2006, p.50) found, were driving up fixed and variable costs. The move to an OA producer pays publishing model, would, Waltham suggests, further increase costs by the need to collect fees from authors and the marketing of services to them. To be financially stable the business model adopted for a society journal will inevitably need to meet the costs incurred. Waltham (2006, p.26) finds that the "OA model as currently construed is unlikely to meet all of these needs" and that there was a "deep concern expressed over the financial sustainability of a switch to this model [OA] across the board". Despite these problems Solomon (2007) has examined at some length the characteristics of five successful OA journals. He found that each manages its finances in different ways ranging from almost entirely unpaid volunteers to one where various income streams have been developed from member subscription, advertising and non-member submission fees.

3.7. Author pays

The DOAJ lists a number of journals published by the Public Library of Science (PLoS), which is a not-for-profit OA journal publisher; its journals are free to the end-user. Its publishing model relies on significant donations from private foundations as well as from the Joint Information Systems Committee (JISC), the Open Society Institute (OSI) and from those authors or institutions who can afford to make some payment towards the cost of publication. (Public Library of Science FAQ… [n.d]). PLoS is typical of the 'author pays' model, where cost recovery has been switched from reader to producer, although total costs may be defrayed by the other sources of funding. The strategy adopted by PLoS has been to establish science-based OA journals in direct competition with high ranking, high impact subscription journals. The strategy has been successful, with seven journals in publication and *PLoS Biology* receiving an Institute for Scientific Information impact factor of 13.9 (Public Library of Science About…[n.d]). The PloS journals are:

PLoS Biology
PLoS Medicine
PLoS Computational Biology
PLoS Genetics
PLoS Pathogens
PLoS One
PLoS Neglected Tropical Diseases

BioMed Central, by contrast, is an independent, not-for-profit publisher committed to OA, making the pledge that all peer-reviewed research articles published by it will be freely available to all and deposited in an accessible repository from the date of publication (BioMed Central OA Charter 2006). BioMed Central funds this publishing model by levying article-processing charges ranging from 0 to $2685 per article. Institutions can pay a membership fee which allows institutional authors to publish freely with them (BioMed Central FAQ…2006). Oxford University Press is a major publisher of journals and books and is a department within Oxford University, thus giving it charitable status, but is committed "to transfers[ing] 30 per cent of its annual post-tax surplus to the rest of the University, with a commitment to a minimum transfer of £12 million per annum" (Oxford University Press 2006). Oxford Journals is part of Oxford University Press and is a not-for-profit publisher, which has developed in parallel with its

existing subscription based service an OA option called Oxford Open (Oxford Journals 2006a) as a further development in its OA programme. This initiative expands Oxford Journal's experiments with OA publishing models. The initiative started in 2003 when *Nucleic Acids Research* moved towards an OA funding model; now the journal is completely OA and is freely available with publishing costs being met by author and institution charges (Oxford Journals 2006b). Oxford Open offers, for a number of its journals, authors or their institutions, the opportunity to pay for their article to be published, assuming it is accepted. If a payment is made, then the article, once published, will be freely available immediately to anyone in an online format. If the author declines to pay then their article, once accepted for publication, will be subject to the normal subscription fees.

This model has been adopted and refined by wholly commercial publishers, who have provided OA on a similar basis by allowing article processing charges to be met by the author, institution or other funders of research. Open Choice offered by Springer (Springer Open Choice …[n.d]) allows, like Oxford Open, authors or their institutions to pay for their article to be published. If it is accepted, again like Oxford Open, the article is posted online for all to access; it will also appear in the printed journal but access will be controlled by subscription; only the online version is freely available. OnlineOpen from Blackwell (Blackwell…[n.d.]) works in almost exactly the same way, offering the same standard of peer-review and formatting as any other article submitted to them, the service was on trial until the end of 2006. Now the option is available to authors to fund their own publication or whose funding agency requires them to archive the final version of their article. For commercial publishers this is a significant concession. In their evidence to the House of Commons Science and Technology Committee (2004, pp.80-81), these publishers believed that the 'author pays' publishing model would compromise peer review. That is, where there was any kind of payment, objectivity in the peer review process would be lost. The assumption was that publication can be bought irrespective of quality. The committee (2004, p.81) acknowledged this concern, recognising that some "lower quality journals" may see all contributions towards their costs as being acceptable. The model of peer review to be adopted by these publishers does seem to be one where the reviewer is blind to the model of payment. Costs are, however, substantially higher than those at BioMed Central, ranging from $1335 to $2670, in some cases dependent on the subscribing status of the individual or the institution (Oxford Journals 2006),

(Blackwell…[n.d.]). As Waltham (2005 p.28) notes, however, a "competitive market is emerging in the level of producer pays fees that publishers are charging to authors".

3.8. Delayed access

HighWire Press is a service hosted by Stanford University which aggregates and publishes the online versions of peer reviewed high impact journals for a range of not-for-profit journal publishers. Of the 4.6m articles available through them, 1.8m are OA (HighWire Press…[n.d.]). This is achieved by the participating publishers specifying a period after which access to their journals is free. Access to journals prior to this period is controlled by payment. This embargo period varies from journal to journal and hence its usefulness can vary; the journal *Brain* published by Oxford Journals, for example, has an embargo period of two years (HighWire Press…[n.d.]). In a similar arrangement, the National Institutes of Health's (NIH) PubMed Central is a digital archive of life science journals which have been deposited by participating publishers. Like the HighWire Press, variable embargo periods are in place (PubMed Central Overview 2005).

3.9. Summary

Authors can make their work OA by either self-archiving it or publishing it in an OA journal. The 'green' self-archiving route requires the author to deposit their work, usually after publication in a toll access journal, in an accessible location on the Internet; this is more frequently being done in OAI-compliant institutional archives. Authors often self-archive both their preprints and postprints. The 'gold' route is where the author has their work published in a journal which is either completely OA or in a conventional journal which has an OA policy that allows authors to publish in them after making a payment. The two types of publishers have different business models. For OA publishers to survive, they need to charge fees to authors or their institutions to have their work published, although some of these publishers may have other sources of funds to help defray the cost of publishing. Toll access publishers rely on their subscription base for the majority of their income but they in turn will charge author-fees if they have an OA option.

Publishers are experimenting at different levels with how best to manage and fund the publishing of OA articles. For small publishers, either conventional or not-for-profit, who

are wholly dependent on a small number of journals for their income this is a difficult process.

Author-fees vary from publisher to publisher. For commercial publishers, fees are generally higher; some publishers will waive author fees if there is a genuine case of hardship. Table 3.1 below summarises some typical costs from 2008.

Table 3.1 Author pays (acceptance) fees for publishing a single article

Publisher	Status	Non-subscriber	Subscriber
Springer	Commercial	$3000	$3000
Blackwell	Commercial	$2600	$2600
PLoS	OA	$1250-2750	$1500
BioMed	OA	$0-2685	$0-2685
Oxford	Not for profit	$2670	$1335

Whilst Table 3.1 gives basic publishing costs for authors wishing to make their article OA, some publishers can offset costs for authors by drawing on other funding. The Table also gives the costs that authors pay if their paper is accepted; no cost will be incurred if the paper is rejected. Some journals charge submission costs irrespective of whether the article will be accepted for publication. The California-based Society for Economic Theory, will, in launching its new OA journal *Theoretical Economics*, be charging a $75 submission fee to all authors ($35 for academics in developing countries) (Lornic 2006).

3.10. Interoperability and the Open Archives Initiative

The possibility, as Harnad (2002) saw it in his original 'Subversive Proposal', was that authors could self-archive their work in digital form to arbitrary websites. This part of his proposal, he realised, was flawed, but it was made before the availability of the:

> Open Archives Initiative (OAI)… Compared to that, both anonymous ftp sites and arbitrary websites are more like common graves, insofar as searching the peer-reviewed literature is concerned. (Harnad 2004b)

Authors may have wanted to self-archive their work, but without being able to have it readily accessed, searched and retrieved, self-archiving would be of little value. The solution to this problem was twofold:

> First the Open Archives Initiative (OAI) created a convention for tagging the critical metadata identifying papers as research articles (author, title, journal, date, abstract, keywords) so that all papers that were compliant with the OAI convention would become "interoperable," meaning that they could be harvested, searched and retrieved as if they were all in one virtual archive containing all and only peer-reviewed research. The second step was to design (free) software that would create OAI-compliant Eprint archives in which authors could immediately deposit all their articles so as to make them openly accessible to all other researchers, thereby maximising their impact. (Harnad 2004a)

Lynch (2001) gives a brief description of the Open Archives Initiative, which appeared in 1999. It did not, he argues, attempt to provide a search tool that could be used to interrogate remote databases; rather, it allowed for a simple method of harvesting the metadata from compliant repositories and storing this into searchable databases.

Free software which uses the OAI protocol is now available: two notable examples are DSpace and Eprints (Pinfield 2003). The software allows users to build an OAI-compliant repository into which documents can be archived; once archived, the metadata from these documents can be harvested using the OAI Protocol for Metadata Harvesting (OAI-PMH) by OAI service providers. Two examples of service providers are OAIster and Citebase, both of which aggregate harvested OAI compliant metadata into a searchable database from which users can directly access documents from their original location (Swan *et al.* 2005, p.26). OAIster (2008) currently has over fifteen million documents which it has harvested from 935 contributors.

3.11. Repositories

There are three basic forms of repository or archive into which OA material can be deposited. These are the subject or disciplinary archive, the institutional repository and the general national repository (Jones, Andrew & MacColl 2006, pp.7-9). Harnad (2002), in support of his ‘Subversive Proposal’, was able to point to the existence of the preprint high energy physics archive set up by Ginsparg as an example of how research could be freely shared. The archive, now called arXiv, was set up in 1991 and was the first notable

large scale example of a disciplinary archive accessible to all (Lamb 2004, pp.144-146). ArXiv is:

> …an automated distribution system for research articles, without the editorial operations associated to peer review… [it was] initiated in 1991, before any physics journals were on-line. Its original intent was not to supplant journals, but to provide equal and uniform global access to prepublication materials. (Ginsparg 2002)

ArXiv has accumulated 466,000 records during its seventeen-year existence and is gathering them at the rate of about 5,000 month (arXiv monthly…2008). The archive currently serves its original users from the field of physics, but also those from other disciplines, most notably mathematics and computer science The archive is used, probably uniquely, by authors both as a repository and as an early publishing tool to quickly publish preprints and receive critical feedback from their peers, which subsequently they will publish in traditional journals to authenticate it to the wider academic community Gunnarsdóttir (2005, pp.565-567). Gunnarsdóttir (2005, p.552) identifies the success of arXiv by quoting Ginsparg, who felt that it was "facilitated by a pre-existing 'preprint culture', in which the irrelevance of refereed journals to ongoing research has long been recognized as irrelevant".

Two distinct forms of scholarly archive or repository are now apparent; the disciplinary type exemplified by arXiv, and those which are "distributed archives located at research based institutions around the world" (Swan *et al.* 2005 p.26).

3.12. Distributed institutional repositories

Research based institutional repositories (IRs) are digital archives into which staff can deposit their work. Ware (2004, p.6) suggests that the rationale for these IRs, as seen by some at least, is as a reform of scholarly communication and publishing and also "to enable the institution to enhance its prestige by making visible the fruits of its faculty's academic and research labours. …[and] are also seen as the digital infrastructure of the modern university". If these IRs are OAI-compliant (and the vast majority are), then the metadata from the peer reviewed research they contain can be harvested and made available to any who wants to access it. The corollary of this process is that free access to

such research will mean greater visibility and impact for its authors and their institutions. This, Crow (2002, p.6) feels, will also allow an IR to concentrate "the intellectual product created by a university's researchers making it easier to demonstrate its scientific, social and financial value".

There are a growing number of institutional repositories that are OAI-PMH-compliant and consequently harvestable by service providers. Currently, the Registry of Open Access Repositories ROAR (2008) has over 1000 repositories registered worldwide, of which 536 are based at research institutions. These 536 archives hold a total of 2,309,512 records, averaging 5087 records each with a median figure of 938. In terms of the two million or so peer reviewed research articles published on a yearly basis, this represents a small but growing part of the total output.

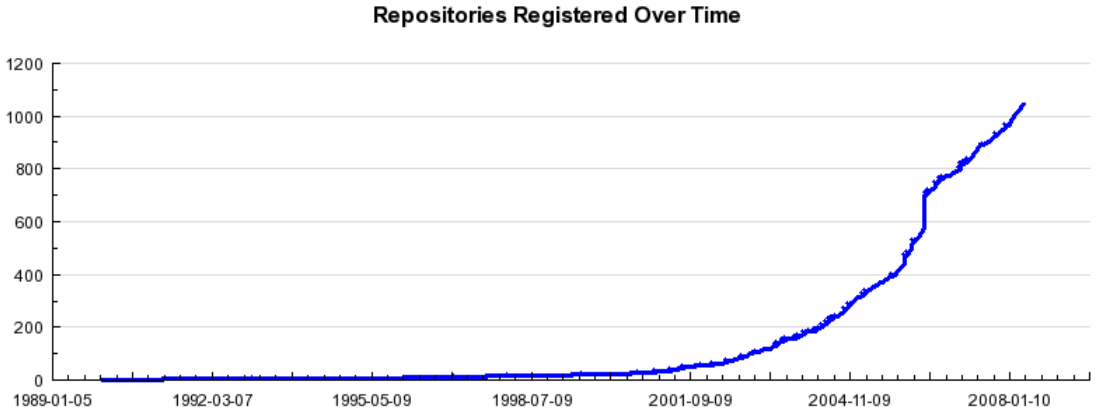


Figure 3.1 Repository numbers (ROAR 2008)

The graph above shows the rapid growth of institutional archives as at March 2008 (Registry of Open Access Repositories (ROAR) 2008). Carr and Brody (2007) caution that growth, however, should be steady rather than spasmodic and that the sustained growth of any archive is more likely to be related to the engagement of the broad community that sustains any archive.

3.13. Self-archiving to repositories

Self-archiving to repositories is not, as Swan and Brown (2005, p.2) point out, an:

> …alternative to publishing in learned journals, but an adjunct, a complementary activity where an author publishes his or her article in whatever journal s/he

> chooses and then simply self-archives a copy. In practice this means depositing the file, which is usually the author's final version of the article after peer review has been completed, in an OA archive or repository.

If all authors adopted this approach, then the majority of peer-reviewed research would be OA, but clearly this is not the case given the above statistics. Filling these repositories with content is probably the biggest challenge for any research institution. As Pinfield (2003, n.p.) remarks, setting up the repository is relatively straightforward but "populating the repository is not".

Whilst the arXiv archive has been very successful in attracting a particular group of scientists to archive their research articles, other authors have not always been so ready to self-archive. Swan & Brown (2005, p.30) found in their survey of OA and self-archiving that the majority (51%) of authors have not self-archived their work to either a disciplinary or an institutional archive, but that this had improved from their earlier survey in 2004 (Swan & Brown 2004b, pp.219-224) when the figure had been 67%. Although these percentages look promising, it is important to note that the overall response rate to the 2005 study was about 5%, i.e., 1296 respondents (Swan & Brown 2005, pp.7-8). Of these 1296, 30% were from an "interested and informed population", although ironically this portion of the survey population had archived the least in terms of the percentage of their work (2005, p.30).

In a large scale survey Rowlands, Nicholas and Huntington (2004, pp.267-268) found that men were more likely to self-publish (on their homepage or departmental website) than women and that the younger the author the more likely it was that they would self-publish their work. The survey also showed that those most likely to self-publish are from computer science, economics and business, mathematics, physics and astronomy, with conference papers and accepted journal papers being the most popular items being self-published. The authors of the survey (2004, pp.268-269) also found that 21% of the respondents had self-archived to an institutional repository, again more men (23%) than women (15%) had done so, and the older the author, the less likely they will have done so.

Harnad, reporting the work of Hajjem (2005), estimated the self-archiving population as being as low as 15%. This figure ties in with the percentage of those archiving their work

to Australian universities; Sale (2005), in his analysis of their 2004-05 deposit rates, finds that the rates are no more than 15% for those universities who have a voluntary self-archiving policy. In contrast, however, the one university that does have a mandatory policy, Queensland University of Technology, has a deposit rate of about 60%. This success is echoed at Southampton University, which has a mandatory departmental policy within its School of Electronics and Computer Science and more recently the whole university; in the School the deposit rate is in the order of 90% (Harnad 2005c). Similarly, the CERN Archive, again with a mandatory OA policy, boasts a collection of 360k full-text documents (CERN Document Server, 2008). Most recently, the faculty of arts and sciences at Harvard University has adopted a self-archiving mandate, which includes the retention of copyright, a policy that is likely to be copied by others. This is the 16th reported institutional or departmental mandate to be adopted worldwide (Harnad 2008). Interestingly, the majority of authors (81%) surveyed by Swan & Brown (2005, p.63) said they would self-archive if required to do so by their employer or research funder.

Ware (2004, pp.17-19) lists the barriers to self-archiving as mainly cultural, where there is general inertia by faculty staff, a lack of awareness and confusion about copyright issues. He also notes that where new institutional repositories have been set up, after the first few months, the deposit rate falls sharply, partly, he reasons, because the potential demand for the service has dried up (2004, p.31). In evidence taken by the House of Commons Science and Technology Committee (2004, p.59), it was found that authors had little incentive to self-archive their work and therefore lacked the motivation to do so, given that publishing in high impact journals is the author's primary aim. The time required to archive documents seems also to be an issue, and has been seen as a significant disincentive for individual self-archiving, although Carr and Harnad (2005, pp.1-5) estimate that it should only take about ten minutes per paper. Swan and Brown, in their 2005 survey, reported a substantial level of ignorance with respect to OA and OA journals, particularly amongst those who had not self-archived at all. Not surprisingly, the most aware were those in disciplines where a self-archiving culture already existed, that is, physics, computing and mathematics (2005, pp.43-46). The knowledge that research by Lawrence (2001, p.521), showed for example, that articles that are available online attract more citations and thus have greater impact than those that are not, seems to have had little effect on authors' perceptions of the benefits of self-archiving.

3.14. Mandatory self-archiving

Given this relatively poor record on self-archiving, it has been suggested by Bosc and Harnad (2005, p.99), Pinfield (2005, pp.30-34), Rowland *et al.* (2004, p.303), amongst many others, that funding bodies should mandate authors to self-archive their work after journal publication. Similarly in their report, the UK's Select Committee on Science and Technology (2004, p.102) recommended that funding bodies supported by government should mandate authors to deposit a copy of all their articles one month after publication, as a condition of their research grant.

The National Institutes of Health in America (NIH) was asked to make a similar recommendation, where it would be a 'requirement' for grantees to make their research findings available online after six months. In September 2004, a proposal was released for consultation where the 'requirement' had been changed to 'request' (National Institutes…[n.d.]). The policy became effective after May 2005. In short, the policy to date has been a failure. The NIH itself provides some statistics on the deposit rate; for the whole of September 2005 this totalled just 275 documents. (NIH Public Access…[n.d.]). Suber (2005c) estimates that the deposit rate should be in the order of 250 manuscripts per day, if all NIH funded researchers were to deposit their research articles. Subsequently in October 2007 the wording was changed to 'require' and was effectively mandating those who receive NIH funding to deposit an electronic version of their peer-reviewed articles no later than 12 months after the official date of publication. This policy became official on January 11th 2008 and becomes effective from April 7th 2008 (Revised Policy on Enhancing Public Access…2008).

Swan *et al.* (2004, p.48) highlight the position in which charities find themselves: having funded research, they find that other researchers cannot widely access the results of that research without subscribing to the journals in which it has been reported. The Wellcome Trust, a charitable organisation and the largest non-governmental funder of biomedical research in the UK, now:

> Requires electronic copies of any research papers that have been accepted for publication in a peer-reviewed journal, and are supported in whole or in part by Wellcome Trust funding, to be deposited into PubMed Central (or UK PubMed Central once established). Note that this requirement will apply to all grants awarded after 1 October 2005, and from 1 October 2006 to all grants regardless

> of award date. [and] Will provide grant holders with additional funding to cover the costs of page processing charges levied by publishers who support the OA model. (Wellcome Trust 2005)

In a similar move during 2005 the UK's Research Councils (RCUK) proposed, subject to consultation, that they too would require research output funded by them to be deposited in an e-print repository (Research Councils UK [n.d.]). By June 2006, these proposals were accepted by most of the research councils such that they required research output funded by them to be archived in a suitable repository; when authors should do this though was not mandated (Research Councils UK [n.d]).

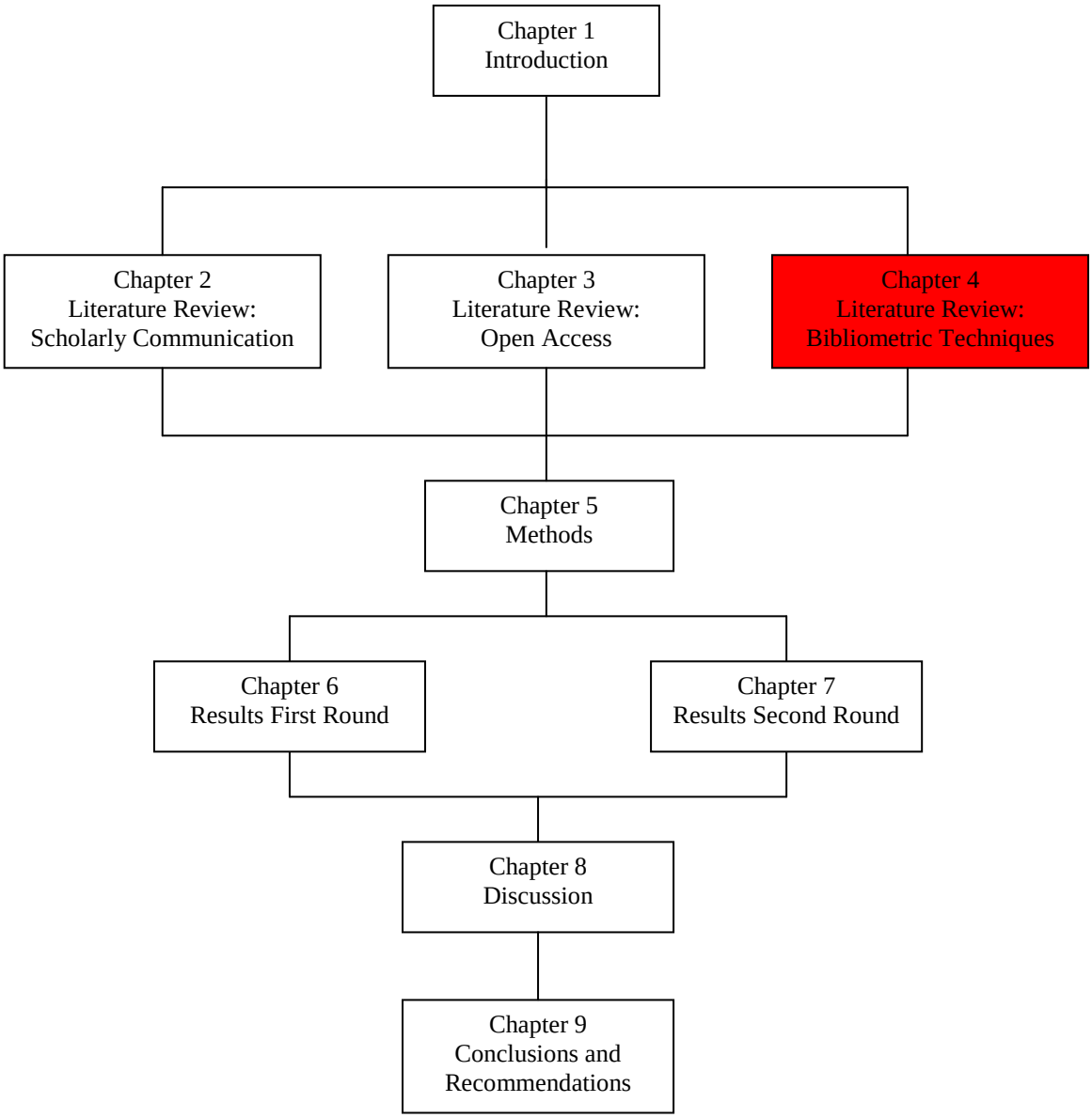
3.15. Mandatory self-archiving and publishers

In the UK, the Association of Learned and Professional Society Publishers (ALPSP) responded to the RCUK proposals; it was concerned that they would have a serious effect on the income of their members (Morris 2005). Morris, aware that libraries are under constant pressure to control costs, thinks that they will increasingly find that "'good enough' versions of a significant proportion of articles in journals are freely available: it is inconceivable that they would not seek to save money by cancelling subscriptions to those journals. As a result these journals would die". This would undermine the work that learned societies fund from any surplus revenue from their journal subscriptions (Morris 2005). Such assertions are strongly rebutted by Berners-Lee *et al.* (2005) who, obliquely at least, point to the findings of Swan and Brown (2005, pp.3-4), who contacted those publishers whom they believed should have been most affected by authors self-archiving their pre and postprint articles to the arXiv repository. Both the American Physical Society and the Institute of Physics Publishing Ltd said they could not identify any losses in subscriptions as a result of the fourteen year existence of arXiv. It might be conjectured that this is also true of the CiteSeer (n.d) repository, which holds, at 2008 some 767,000 documents primarily related to computer science. Berners-Lee *et al.* (2005) rebut the claims of Morris on the basis that journal publishers are losing, or have lost, subscription revenue by making entire journals OA along with additional functionality, rather than by authors self-archiving single articles to a repository.

Oppenheim (2005, p.6), one of the signatories to the above rebuttal, in an earlier editorial which considered OA in the light of the UK Science and Technology Committee Report,

viewed the impact of repositories as significant. He concluded his summary with the prediction that “many smaller scholarly publishers (including, I fear scholarly societies) will suffer because of the competition from the repositories, but the larger, more innovative ones will continue, and will continue to be profitable”.

Chapter 4 Literature Review: Bibliometric Techniques



4.1. Bibliometrics

There have been a number of studies which attempt to demonstrate that authors who make their academic articles freely available can achieve greater impact with their work than those whose work is not. These studies use citation analysis to justify their results. Authors may cite the work of others in articles which report their work; when they do so they are creating a bibliographic record that can be counted. Once aggregated, these records and the links between them can be subjected to statistical analysis from which conclusions may be drawn. Citation analysis is, however, only one technique amongst a family of techniques that are found within the field of bibliometrics. Diodato (1994, p.13) in his *Dictionary of Bibliometrics* gives eight definitions of the term bibliometrics, the first of which comes from F.W. Lancaster, who defines bibliometrics as "The application of various statistical analyses to study patterns of authorship, publication and literature use…"

Historically, bibliometric studies have fallen, initially at least, into two broad categories (Borgman & Furner, 2002, p.8). In the first category are those studies that evaluate and often assess the suitability of the tools and theories used in bibliometrics. The second type are those studies which "use bibliometric methods in order to describe, explain, predict, and evaluate the communication behavior of scholars" (Borgman & Furner, 2002, p.8).

The second type itself may be divided into two distinct types. Borgman and Furner (2002, p.11) describe 'Relational Link Analysis', where:

> Link counts are used as indicators of the level of connectedness, the strength of relationship or the direction of flow, between documents, people, journals, groups, organizations, domains or nations. Relational citation analysis is used to answer research questions of the type, "Who is related to whom?"

The authors (2002, pp.11-12) then consider 'Evaluative Link Analysis', where link counts are used:

> …as indicators or measurements of the level of quality, importance, influence, or performance, of individual documents, people, journals, groups domains (subject areas, fields, or disciplines) or nations. Evaluative link analysis is used

> to answer research questions of the type, "Whose research or influence is better, or has greater impact, than whose?

Moed (2005, p.x) describes evaluative bibliometrics as:

> …a subfield of quantitative science and technology studies, aimed to construct indicators of research performance from a quantitative analysis of scholarly documents. Citation analysis is one of its key methodologies.

There have been many hundreds of studies which have attempted to evaluate the performance of academics, their departments, their institutions and even the country in which they work by using evaluative bibliometrics tools.

4.2. History of bibliometrics

Pritchard questioned, in his brief 1969 note Statistical Bibliography or Bibliometrics? (1969, pp.348-349), which of the two was the more appropriate term for this field of study; he settled with coining the term 'bibliometrics'. He thought it should:

> …shed light on the processes of written communication and of the nature and course of development of a discipline (in so far as this is displayed through written communication) by means of counting and analysing the various facets of written communications. (Pritchard 1969, p.348)

Bibliometric methods are concerned with the study of the formal channels of scholarly communication (Borgman, 1990, p.14). This general field of study has expanded to embrace new terms which in themselves seek to incorporate the diversity of scholarly communication. Egghe (2005, p.1311) uses the broader term Informetrics to "cover all-metrics studies related to information science…scientometrics (social policy, citation analysis, research evaluation…) webometrics (metrics of the web., the Internet or other social networks…)".

The analysis of these formal channels of scholarly communication by standardised metrics is essentially a twentieth century device, with its origins in the study of bibliographical records. The few early studies were principally interested in the statistical analysis of library collections and the nature of particular disciplines (Hertzel 2003, pp.292-294). The first notable study was by Cole and Eales in 1917, who examined the

history of animal anatomy through 6436 publications taken from the period 1543-1860. They analysed the data by country and subject and could demonstrate the influence of individuals, events and public bodies on the research as reflected in the subject matter within the literature current at the time. Hertzel (2003, pp. 296-297) gives a second important example related to growth of scientific literature, where the contents of 13 annual issues of *The International Catalogue of Scientific Literature* were analysed by Hulme. Hulme was able to show the productivity of different countries and the way various subjects grew or decayed depending on the way the subject was influenced by general changes in the external environment. Interest in the subject grew, with an increasing number of studies looking at a larger range of subjects. Meadows (2000, p.88) has quantified this growth, noting an increase in the number of bibliometrics studies rising from 10 in the first decade of the twentieth century to over 350 by the sixth.

4.3. Laws of bibliometrics

Three notable bibliometric 'laws' were developed in the twentieth century. These have helped scholars understand author productivity and the way the literature is scattered within a particular discipline. They are associated with the authors who first derived them. Lotka was concerned with author productivity, Bradford with the way articles are scattered and Zipf with word frequency. These distributions are amongst a family of inverse square power laws which can be used to model a range of naturally occurring phenomena: Newman (2005, p.323) notes their occurrence in physics, economics, finance and other fields. These three 'laws' are probably the most well known in bibliometrics; however there are many others: Diodato (1994, p.99) lists another seven, quite apart from the many mathematical variants that have refined and developed these laws.

The three distributions have been used in a variety of applications; these have ranged from determining which authors are most productive in a particular discipline to identifying how library collections can be managed to ensure the core literature in a given subject is identified and provided. Bence and Oppenheim (2004, p.470) examined the scatter of business and management articles submitted for the UK's Research Assessment Exercise and found a Bradford-Zipf distribution. Rowlands (2004, pp.5-10) examined author productivity in terms of how likely it is that an author will return to a particular publisher to publish their next work. Using Lotka's law, he found that author data

provided by Emerald Group Publishers fitted the distribution very closely. A simple title search, incorporating the names of Bradford, Lotka and Zipf, of the database LISA for the last ten years returned 55 journal articles, suggesting that there is still significant interest in these bibliographic distributions.

4.4. Lotka's Law

Lotka collected and analysed the authors referenced in the Chemical Abstracts Index for the period 1907-1916. He recorded only those authors whose surnames began with A or B. All the names were collected from *Auberbach's Geschichtstafeln der Physik* from its inception to 1900. Lotka listed the authors and the number of articles they had written; from this he was able to derive a relationship between authors and the number of articles they had written (Lotka 1987, p.113). From these observations Lotka formulated a general equation:

$$x^n y = c$$

where y is the portion of the authors making x contributions each and where n and c are the parameters that depend on the field being analysed (Diodato 1994, pp.105-108). As Lotka (1987, p.119) explains:

> In the cases examined it is found that the number of persons making 2 contributions is about one-fourth of those making one; the number making 3 contributions is about one-ninth, etc.; the number making n contributions is about $1/n^2$ of those making one; and the proportion, of all contributors, that make a single contribution, is about 60 per cent.

Table 4.1 illustrates this:

Table 4.1 Illustration of Lotka's law – author productivity (Oppenheim [n.d])[1].

Number of papers published (n)	Number of authors publishing (n) papers
1	100
2	25
3	11
4	6
5	4
…	…
100	1

In essence, many authors will produce one or two articles but as author productivity rises there is a proportional drop in the number of authors producing articles.

4.5. Bradford's Law

Bradford investigated the way articles on a given topic were scattered throughout journals. Bradford examined two fields; applied geophysics and lubrication. By consulting bibliographies in these two fields, he was able to compile a list of the journals and their productivity by how many relevant articles they contained (Bradford 1948, reprinted in Meadows 1987, pp.148-149). From these data Bradford observed that the journals could be separated into three groups, the first group containing journals which were the most productive in terms of relevant articles and the second and third progressively less so. This led him to state that:

> …if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups of zones containing the same number of articles as the nucleus, when the number of periodicals in the nucleus and succeeding zones will be as 1: n: n^2 (Bradford 1987, p.152).

[1] Oppenheim, C., [n.d.]. *Bibliometrics research methods*.

In the case of Bradford's (Bradford 1948, reprinted in Meadows 1987, p.150) original work on lubrication, he found from 164 journals 395 articles. These he grouped:

8 journals produced 110 articles;
29 journals produce 133 articles;
127 journals produce 152 articles.

This gives eight core journals and a relationship of roughly 4, i.e., 8, 4*8, 4^2*8, giving a theoretical distribution of 8, 32, 128 journals; not so dissimilar from the actual data he found.

4.6. Zipf's Law

Zipf was concerned with the frequency with which words occurred in a given text. He was able to derive a relationship between the size of a word in terms of the number of letters it contained and its occurrence. Initially he concluded that:

> If the number of different words occurring once in a given sample is taken as x, the number of different words occurring twice, three times, four times, n times in the same sample, is respectively $1/2^2$, $1/3^2$, $1/4^2$,…$1/n^2$, of x up to, though not including, the few most frequently used words; that is, we find an unmistakable progression according to the inverse square, valid for well over 95% of all the different words used in the sample. (Zipf 1965, p.xii)

Meadows (1998, p.121) explains this: "For example, a word ranked tenth in terms of frequency is used one-tenth as often as the word ranked first [and] the words that appear most frequently are also, on average, the shortest words". Zipf developed the formula $ab^2 = k$ to describe his law, where a represents the number of words of a given occurrence and b the number of occurrences (Hertzel 2003, p.312).

4.7. Cautions and convergence of the laws

Hertzel (2003, p.303) cautions that the above three 'laws' are not true laws in the sense of laws that can be proved by consistent demonstration or by a mathematical proof to give repeatable results when using different data. Later results using Lotka's law have shown it to be often at variance with his original work, the most notable difference being in the distribution of single authored works. Thus different exponents have been used ($1/n^2$,

$1/n^3$, $1/n^4$ etc) in the calculations to get a closer match to the empirical data. Price (1963, pp.46-48) has also suggested that Lotka's law needs modification for the case of those highly productive authors. Hertzel (2003. pp.304-305) identified a number of disagreements about the applicability of Lotka's law and its careless application. Hertzel (2003, p.305) concludes in her discussion of Lotka's law by quoting Pao, who acutely observed that "…studies have assumed the inverse square relation as the basis for testing. Others derived the value of the constant c from the percentage of single paper contributions. None of these assumptions can be traced back to Lotka". Neither have Bradford or Zipf escaped careful examination and critical scrutiny. Vickery (1948, p.198) found inconsistencies in the way Bradford's law was stated and its algebraic expression; the algebraic expression Vickery corrected such that the two were consistent. A number of authors have concluded that the 'laws' of Bradford and Zipf are essentially the same thing (Kendal 1960, p.35; Leimkuhler 1967, pp.197-207). Leimkuhler (1967, p.207) considers they are "just two different ways of looking at the same thing". Brookes (1969, pp.58-60) summarised "the outcome of recent analyses of empirical data which have enabled the general form of the Bradford-Zipf distribution to be elucidated [and that its form] has two components... an initial rising curve…running smoothly into…a straight line…". Bookstein (1976, p.416) has argued that:

> One of the more surprising findings in the information sciences is the recurrence of a small number of frequency distributions. [and] a point of view is adopted that allows us to understand them as different versions of a single distribution.

Bookstein (1976, p.419) agrees with other writers who have shown that "Lotka's law and Bradford's law are different approximate descriptions of the same basic distribution, and that at least for larger values of r both agree with Zipf…the distributions are approximately the same; it is only the entities and events that differ".

4.8. Citation analysis

When writing an academic paper or monograph, scholars will nearly always support their work by making reference to the research of others, by citing them. In this referencing process authors will record at the end of their document the bibliographic details of the publication they have referred to. Smith (1981, p.83) differentiates the term 'reference' from 'citation': "A reference is the acknowledgement that one document gives to another:

a citation is the acknowledgement that one document receives from another". Hertzel (2003, p.317) suggests that a "relationship is implied between the cited document and the citing" and Smith (1981, p.83) continues by explaining that "Citation analysis is that area of bibliometrics which deals with the study of these relationships".

Citation analysis, like bibliometrics in general, has its particular history. The technique has been in use since the early 20th century, however its application has now become commonplace. Citations are, Garfield (1979, p.1) explains, the:

> …formal, explicit linkages between papers that have particular points in common. A citation index is built around these linkages. It lists publications that have been cited and identifies the sources of the citations.

In essence, in its most basic form, citation analysis simply counts the number of citations that an article or monograph has received in a particular period. For every citation counted, there will be the associated bibliographic information of the citing and cited document. These bibliographic elements of, for example, author(s), discipline, nationality, country, affiliation and document types can become metrics themselves, and the frequency of their occurrence and the relationships between them can be used in bibliographic analysis (Borgman & Furner 2002, p.7). In evaluative studies, the most fundamental measure is that of citedness, that is, the number of times an individual is cited or perhaps the number of times a particular article is cited, or the number of times the content of a particular journal is cited (Borgman & Furner 2002, pp.12-17).

To be effective, the counting of citations requires a database of citations where the links between citing and cited documents can be readily identified, counted and retrieved. This process was not easily undertaken until Eugene Garfield launched the *Science Citation Index*. The first annual edition was published in 1963 (Garfield 1979, p.xiii) followed later by siblings *Social Sciences Citation Index* and the *Art and Humanities Citation Index*. Garfield (1979, p.7) acknowledges the influence of Shepard's Citations: this was the first major citation index which was regularly updated; it dates from 1873 and was used to search for legal precedents. Garfield was working at the same time on the Welch Project, a project sponsored by the Armed Forces Medical Library which was examining subject indexes to the medical literature. While the project team was examining review articles, Garfield realised that every sentence in a review was supported by a citation to a

previous work and as such "He realised that a review article could really be considered as a series of indexing statements" (Baird & Oppenheim 1994, p.4). Connecting the format of Shepard's Citations and the ideas generated from the Welch Project, the "format of a citation index for scientific literature was developed" (Baird & Oppenheim 1994, p.4). Garfield published his theoretical ideas in 1955 "in which a citation index to the literature of science was visualised as following the model of *Shepard's Citations*" (Garfield 1979, p.10.). Hertzel (2003, p.319), quoting Garfield, describes that:

> A citation index is an ordered list of cited articles each of which is accompanied by a list of citing articles. The citing article is identified by a source citation, the cited article by a source citation. The index is arranged by reference citations. Any source citation may subsequently become a reference citation.

Garfield was writing the above before the indexes were available electronically, but the interconnections between the cited and citing articles remains the same. Data from the three indexes have been used over the last four decades or so for many hundreds of studies (Moed 2005, p.13). These citation indexes are now owned and operated by the commercial company Thomson Reuters.

4.9. Journal coverage and impact factor

Of the estimated 26,000 peer reviewed journals currently in print, the ISI indexes approximately 8700: these ISI call 'core journals'. The indexing process extracts from each journal all of the bibliographic data including all of the cited references (The Thomson Scientific…[n.d.]). These 8700 journals are classed as being the most prestigious journals in their respective fields (*Web of Science* [n.d.]). Thomson ISI makes no attempt to index all peer reviewed journals; as justification, it points to Bradford's Law on scatter. The 'law', as discussed above, has demonstrated that a relatively small number of journals publish the bulk of significant scientific results; on this basis, journals are selected and deselected on a rolling basis depending, amongst other factors, on their coverage and usefulness (The Thomson Scientific…[n.d.]). In this selection process, citation statistics are also considered, including overall citation rate, journal impact factor and immediacy index.

A journal's impact factor is calculated from the number of citations that a journal has received over a two year period: "Thus, the impact factor of a journal is calculated by

dividing the number of current year citations to the source items published in that journal during the previous two years" (The ISI Impact Factor [n.d.]). From these impact factors, journals are ranked within their subject disciplines on an annual basis. These rankings tacitly confer status and prestige to those journals which are highly ranked; they become the journals in which authors seek to have their articles published. Different disciplines vary in their citation rates and the point at which the average article in its discipline is most heavily cited. Thus, different disciplines show different impact factors (Moed 2005, pp.94-95). Ranking journals merely by their impact factors across disciplines is therefore misleading and inappropriate (Moed. 2005, p.95).

4.10. The meaning of citations

These indexes were designed primarily for the retrieval and dissemination of scholarly literature, but, as Garfield soon realised, such bibliographic databases could be used to understand the nature of scholarly activity. Garfield's (1979, p.62) view was that:

> If the literature of science reflects the activities of science, a comprehensive, multidisciplinary citation index can provide an interesting view of these activities. This view can shed some useful light on both the structure of science and the process of scientific development.

Many authors have shed light on the structure of science by using the citation counts found in the three indexes and many of these have been evaluative in nature, that is, they have been used, for example, as indicators of quality, importance or influence (Borgman & Furner 2002, p.11).

4.11. Why people cite

Baird and Oppenheim (1994, p.3) consider that people cite the works of others because it is relevant to the argument they are making. In doing so, Baird and Oppenheim assert, an "author may be criticising the earlier item, may be building on it, may be using it to enhance his or her argument, or it may be acknowledging an early pioneer". The actual reasons why people cite a particular work rather than another are many: Smith (1993, pp.375-376) lists fifteen from Garfield; Baird and Oppenheim (1994, p.6) list seventeen. Taking a selection from these two, the reasons include:

- paying homage to pioneers in a field;
- giving credit to related work;
- providing broad background to the topic;
- correcting or criticising the previous paper;
- citing a major figure because you think he or she may be a referee of the paper when you submit it to the journal;
- citing a major figure because it makes your research look more respectable;
- identifying methodology, equipment, etc;
- alerting the reader to forthcoming work;
- disclaiming work or ideas of others (negative claims);
- disputing priority claims of others (negative homage).

Citation studies are often criticised for treating citations of being of equal value, when clearly people's motivations for citing can vary so markedly and the range of the material from which they have chosen to cite can be very broad or very limited (Smith 1993, pp.84-85). Further complications arise when the citation patterns of different disciplines are examined. Different disciplines have different citing behaviour and coverage within the ISI databases; this has led to the observation that there are discernible differences between science, social science and the humanities when citation patterns are examined (Moed 2005, pp.147-151). Cronin (1984, p.16) thinks that scientists " cannot realistically claim to work in a social vacuum", the implication being that to understand the motivations of scientists it is necessary to think of them working in a closed community but as part of a social system. This social system, it is argued, is evident in the work of White (2001, p.87), who looked at the tendency of authors "to recite themselves and others in multiple works over time". In this research, White looked at the citation and recitation behaviour of eight information scientists and found that most "of the eight are affected by social networks – that is they cite authors whom they know personally from school, the workplace or an invisible college" (White 2001, p.93). Cronin (2005, pp.84-85) confirms this with an analysis of the citing behaviour of three authors: he found that their citing behaviour was closely linked to people they knew and in that turn there was a similar relationship with those that had cited them.

Studies that have counted citations, and then have drawn some conclusions perhaps about the quality of the journals from which the citations were counted, about their authors or

about the institutions where the authors work are often viewed as being controversial. This is especially true if related to these conclusions financial rewards and individual status are attributed. The questions frequently asked are, what do citation counts measure, and why do people cite particular works as opposed to others? Moed (2005, p.193) considers that:

> Citations are manifestations of underlying processes that may be studied from various disciplinary perspectives [and that] to understand what citations indicate, and to relate citation counts to common concepts in evaluative bibliometrics such as research performance, 'scholarly quality', 'influence' or 'impact' insight is needed into the nature of such processes. Their theoretical understanding contributes to what is often denoted as a 'theory of citation'.

This said, there seems to be a range of views about what constitutes a valid and consistent viewpoint. Moed (2005, pp.193-198) suggests that there are in fact five main disciplinary viewpoints from which citation indicators are constructed, used, interpreted or theoretically founded.

Physical approach	This approach is exemplified by Price, where the measurement and quantification of scholarly activity would lead to the construction of suitable indicators.
Sociological approaches	Deal with scholars' perceptions and conceive both scholars' statements and their publication and referencing behaviour as social acts.
Psychological approaches	Studies in this area typically analyse citer motivations.
Historical approaches	Look at the historical development of ideas thorough citation analysis. Garfield's work on historiography typifies this approach. There have also been studies on the social, economic and institutional conditions under which research is carried out.

Information and communication scientific approaches	Studies in this approach are concerned with the study of scholarly communication. Borgman typifies this approach.

Authors writing in this area may of course adopt any of the above approaches, depending on their particular area of interest, the subject they are examining and the results they are hoping to find.

4.12. Issues with citation analysis

Quite apart from the viewpoint from which any bibliometric study may be conducted, many authors have discussed issues and problems associated with citation analysis, some of which are noted here (Smith 1981, pp.87-93: Baird & Oppenheim 1994, pp.7-9: MacRoberts & MacRoberts 1996, pp.436-438: Liu 1993, pp.376-378).

It cannot be assumed that if an author has cited someone they have read the article or that each article cited is of equal influence and value to the author's argument. Authors will often take only a small part of an article and cite it in support of their own work (Smith 1981, pp.87-93; Baird & Oppenheim 1994, pp.7-9). Liu (1993, p.377) reports the research of Moed and Vriens, who found that errors in the metadata of reference citations were often repeated by other authors, who presumably had copied the reference incorrectly, possibly without having read the article. This is quite apart from the actual citing errors which authors routinely make (Baird and Oppenheim (1994, p.7). Moed (2005, p.49) found from an exhaustive comparison of 22 million cited references matched to 18 million target articles that 7% had errors in the bibliographic reference to the cited article. Smith (1981, p.88) notes that authors do not always cite the best possible works, partly because they may not have access to all or some of those works. MacRoberts & MacRoberts (1996, p.436) carefully scrutinised fifteen papers to see if all the influences evident in the papers had been cited. They found that where the authors should have collectively cited 719 references, they had only made 216 citations, thus clearly citing only a fraction of their influences. MacRoberts & MacRoberts (1996, p.436) also found from another sample of writers evidence of bias in the citations, where authors had sometimes consistently cited particular facts whilst completely ignoring other citable facts; in addition, authors often took their citation from a secondary source.

White (2001, p.93) studied the citing behaviour of eight information scientists, and showed that all eight cited themselves more frequently than any other author. Snyder and Bonzi (1998, p.431) found that self-citation rates across a number of disciplines was on average 9% with a range of 3-15%; rates within disciplines remained fairly consistent. MacRoberts and MacRoberts (1996, p.441) were very cautious about the value of scientific papers and the citing behaviour which accompanies it; they concluded that "The formal paper presents a story, a nice rational story, but not the story, and the citations present an array, but not the only array possible".

4.13. Validity of citation studies

It is clear from the above that citation studies which simply count the number of citations that authors have received and then rank their performance on their citation count alone can be misleading. Similarly, simply using the impact factor of the journals in which an author has published as a guide to the quality of their work can also mislead (Moed 2005, p.29). Seglen (1992, pp.632-633) has noted that the distribution of citations that a journal receives will be generally highly skewed, with a few articles generating the majority of the citations and thus disproportionately affecting journal impact factors. Even more of a concern for those writing in academic journals, who might be measured on their citation count, is that in many cases they are not cited at all; in one pilot study reported below 44% of the articles were not cited, and in many cases these articles had multiple authors. Seglen (1992, p.629, 635) reported levels of uncitedness of 50%, with larger variations occurring within different disciplines; Garfield (1979, p.240) found 25% of scientific articles to be uncited. Garfield (2005, p.8) has also noted that from an analysis of 38 million articles covering the period 1900-2005 about half had not been cited at all.

As Moed (2005, p.37) notes, "citations measure impact rather than quality: measuring and valuing citation impact are analytically distinct". Rather, citation analysis provides evidence of impact and this evidence has to be carefully weighed and judged within the limitations of the data from which it has been drawn. Despite these criticisms, citation analysis has been used to assess the performance of journals, individuals, groups, university departments and the productivity of nations (Borgman & Furner 2002, pp.14-21). Central to the use of citation analysis is the hypothesis posed by Borgman and Furner (2002, p.24) that:

> The validity of using citation counts in evaluative citation analysis rests on the truth of a hypothesis that does more than suggest simply that the probability of a citation of an earlier document by a later one varies partly with the level of quality of the earlier document and partly with other variables: it proposes that the quality of the earlier document is the most significant factor affecting its citation count.

Baird and Oppenheim (1994, p.8) list the areas where citation analysis has been used to compare citation counts to other objective or subjective studies, for example of the prestige of a university within a particular country. They conclude (1994, p.8) that, despite the 'noise' and the many valid criticisms, citation analysis is the "fact whatever measure you take for the eminence of an individual scientist or of a journal or of an institution, citation counts provide strong correlation with that result". Such correlations have been evident in the Research Assessment Exercise (RAE) in the UK. The RAE ranks, principally by peer-review, the quality of the research produced by the authors in the participating university departments, and from this ranking, research funding decisions are made. Oppenheim, over a series of articles (1995, pp.18-27: 1997, pp.477-487: Norris & Oppenheim 2003, pp.709-730), compared the ranking of university departments (Library and Information Science, Genetics, Anatomy and Archaeology) with the collective citation counts from the authors involved. In all cases the correlation between ranking and citation counts was high and statistically significant. A study by Smith and Eysenk (2002, pp.6-7) which examined psychology departments found a similar result.

Moed (2002, pp.731-732) lists a number of key questions that should be asked of anyone producing bibliometric statistics, including, amongst others, the verification and validity of the sources used, the coverage of the particular subject in the ISI databases, the way in which the publication data was collected and ensuring that those being evaluated are properly consulted. He concludes that:

> Bibliometric indicators reflect scientific impact not quality, and provide useful supplementary tools in the evaluation of academic research, provided that they have a sufficiently high level of sophistication: that their pitfalls are taken into account; and that they are combined with more qualitative types of information.

4.14. OA and research impact

As discussed earlier, Harnad (2005a) has calculated that £1.5bn may be being lost in research impact because researchers cannot access all the research they would like, due to the financial barriers erected around the publications in which this research appears. The general consensus is that this problem can perhaps be most easily resolved by authors self-archiving their peer reviewed work to an institutional or disciplinary archive where it can be found by anyone. Few authors have, however, done so (Swan & Brown 2005, pp.62-68). When Lawrence in 2001 carried out his study on the comparative citation advantage of freely available conference articles against those that were not, and found that there was a significant citation advantage to those authors whose work was freely available, he was making a powerful and reasoned argument for self-archiving. Given that the number of citations received by an author is a proxy measure for the quality of their work, then it seems logical that research impact is increased by OA (Antelman 2004 pp.372-374). It logically follows that if everybody self-archived their work, the advantage would disappear.

There are a number of studies which directly address the question of whether OA articles receive more citations than TA ones, or that online articles receive more citations than offline ones. These studies fall into three particular types:

- single disciplines, such as computer science or astronomy;
- specific journals where the contents alone are evaluated;
- multiple fields where a range of subject disciplines are examined.

In addition, there are a number of articles, which address issues that can be related to OA in a broader context, many of which are concerned with trying to establish what the causes of any OA citation advantage may be. These studies are discussed in section 4.18.

4.15. Research impact in discrete disciplines

Lawrence's study, starts from the premise that "usage increases when access is more convenient" (2001, p.521). Lawrence makes the point that computer science was a forerunner in making research available on the web and its output was readily accessible through search engines such as *Google*. Lawrence extracted approximately 120k conference articles from the Digital Bibliography & Library Project (DBLP) (Computer

Science Bibliography 2006), a bibliographic database of computer science articles and proceedings. Since DBLP contains only the metadata to each conference article, it is assumed from his study that Lawrence matched these records with those found in ResearchIndex (now CiteSeer) (CiteSeer 2006), since Citeseer give citation counts and the functionality to assess their OA status. From this data, Lawrence estimated online availability and citation counts and showed that the probability that an article is OA can be clearly correlated to the number of citations an article had received and the year of publication from which the article was made freely available (2001, p.521-522). He concludes that the mean number of citations to offline articles is 2.74 and to online articles 7.03; taken as an average across all the years analysed, online articles were cited 4.5 times more frequently than offline articles. Similarly an analysis of off and online articles at different conference venues showed that 90% of the online articles by conference venue were more highly cited. He also concluded "More highly cited articles, and more recent articles are more likely to be online" (Lawrence 2001, pp.521-522).

Lawrence did not give the number of OA articles found, nor how these were distributed among authors, nor whether citation counts were clustered around a notable group of authors or journal. Neither was it possible to discern what type of publications these were, or whether, despite being conference articles, they later appeared in academic peer-reviewed journals. Such information would have been useful in determining whether the data was concentrated amongst a relatively small number of authors or authors with very high citation counts, or in fact whether there was a preponderance of very high or very low impact journals. There is evidence from Lawrence's work that very prestigious conferences produced articles with noticeably higher citation counts than their lesser rivals. The final conclusion drawn by Lawrence is that online articles are more highly cited than offline ones, although he did not compare directly the citedness of authors from the same conference who did or did not make their articles available online. Nevertheless, given the magnitude in the increase in the number of citation counts, these results are significant, and Lawrence's article itself has been heavily cited since its first publication.

Kurtz (2004) discussed whether a restrictive access policy to journals reduces the frequency with which readers can access the full text of articles. He compared the online access rate to a number of physics and astronomy journals; there was a clear relationship between those who had accessed full text and the journal publisher's access policy. The

more restrictive the access policy, the less frequently full text articles were accessed. In an extension to this earlier work, Kurtz *et al.* (2005, pp.1396-1397) address the issue of whether greater access brings greater research impact within the field of astrophysics. They examined three possible causes of this effect that greater access gives a higher citation count, that early pre-print publishing and access through arXiv gives longer exposure and higher citation counts and that the self-archiving practises by authors of, for example, their most important and citable papers increases citation.

Their results show that increasing access to existing articles did not increase the probability that they will be cited. This finding is not in agreement with other studies that have attempted to find such a relationship. In justification of this finding, Kurtz *et al.* (2005, p.1401) consider that anyone publishing in astrophysics must already have access to the core journals in the discipline to be able to publish, and so do not need an alternative OA route to the relevant material. Publishing articles early through arXiv moves the peak of citations relative to their age to an earlier point in the article's citeable life. Brody, Harnad & Carr (2005), who also found a similar effect, confirm this finding. Kurtz *et al.* (2005, pp.1398-1401) also considered that some sort of selection process based on quality must be occurring when authors self-archive to arXiv, given that only six out of the top 200 most cited articles had not been archived to arXiv. The distribution model they used had predicted that 16 non-arXiv articles should have appeared in the top 200 most cited articles if there had been no bias in the way authors had selected their work for archiving (Kurtz *et al.* 2005, pp.1398-1401).

Harnad and Brody (2004) report the results of an analysis of the physics articles indexed by the Institute of Scientific Information between the years 1992-2001. They compared the citation count of those articles that had been self-archived (making them openly accessible) by their authors to the citation counts to those authors who had not. They were able to show that there was a significant advantage in terms of the number of citations received by a self-archived article, ranging from 2.5 – 5.8. In this comparison, the authors established the status of the articles by reference to the arXiv database. As discussed earlier, this database contains significant pre and postprint articles from the physics research community, so it is less surprising that there is a citation advantage. Overall, the percentage of OA articles to Non-OA articles was 10%. Like the work of Lawrence, it is not possible to tell from the research whether very high individual citation

counts skewed the results found or whether there was a group of consistent self-archiving authors. Additionally, the point at which those articles that were made OA was not given, so they may have had a longer period in which to accrue citations when compared to their TA counterparts.

4.16. Research impact within particular journals

Anderson *et al.* (2001) did not reach the same conclusions. Their work focussed on several measures of a freely available online section of the print/online journal *Pediatrics* (Pediatrics electronic pages). This online only section was launched in 1997 to allow the publication of ten to fourteen articles every month; the articles were peer reviewed and of equal quality as those appearing in the print/online version. However, it was the case that articles were selected by the editor for the online section, therefore possibly adding some editorial bias in their selection. Three measurements were considered for both versions of the journal for the period 1997-1999:

- web usage statistics;
- citations within the biomedical literature;
- author perceptions.

Not surprisingly, the web usage statistics were higher for the online section of the journal, since this was the only way it could be accessed. Overall, PDF downloads were over three times higher for the online version compared to the print/online only version, and on a per article basis, the online downloads were six times higher. The rate of decay in access was similar for both journals. Citation counts for the online only section were consistently below that of the print/online version, with a mean difference of 3.09 ± 0.93. Similarly, far more of the online only articles did not receive any citations at all. The authors also conducted a reader survey, which showed a preference for print and the printed version of the journal. The survey also found that authors viewed online publication as being "second-tier" to print publications, although a large percentage of authors had included these on their CV (Anderson *et al.* 2001).

The work of Anderson *et al.* (2001) has only recently been widely read, although it was cited by Odlyzko in 2002. In the light of more recent studies, their results look to be out of place. However, the period they examined, 1997-1999, was well before the OA movement was evident in terms of international declarations and the widespread

availability of repositories. Compared to the studies carried out by Lawrence (2001) and by Harnad & Brody (2004), the Anderson study is small. It is also relevant to note that the comparison was between one set of authored articles and another under the umbrella of a single title which, for the majority of its articles, was subscription only. Access is also an issue: whether the content of the online version was widely indexed and accessible via web search engines may have restricted its exposure to online searches. Access is clearly a precondition to being cited. Harnad (2005e) criticised the study for its small scale and the fact that it was comparing print versus non-print, not OA versus non-OA. That is, not comparing the citation counts of authors from within the same journal where some of its authors have self-archived their work and others have not. Despite this valid criticism, the research of Anderson and his colleagues was not that dissimilar to the methodology adopted by Harnad & Brody (2004) in that the best way to test the impact advantage of OA was to "compare the citation counts of individual OA and non-OA articles appearing in the same (non-OA) journals".

Wren (2005) carried out research within the biomedicine field on 13 toll and four OA journals by the interrogation of Medline records and the use of *Google* to discover, amongst several other objectives, "to what degree OA publications are shared on non-journal websites". He found a significant correlation between the impact factor of the journal and the frequency of occasions on which an OA online version of an article from the same journal could be found. Although not new, this finding relates to the work of Kurtz *et al.* (2005, p.1400), discussed above, where there is a suggestion that the more highly rated authors may tend to self-archive more frequently than other authors are, an effect he dubbed the 'trophy effect'. Antelman (2006a) has noted, however, that Wren did not actually know the source of the open access copies he found, so the extent of the 'trophy effect' cannot be wholly assessed from his work. Antelman, using data she collected, was able to show that from three of the journals that Wren looked at and in particular, the *New England Journal of Medicine*, only 12% of those articles that were made OA were posted by their authors.

Eysenbach (2006) took a selection of articles that appeared in a single journal (*Proceedings of the National Academy of Sciences*), some of which were TA and others which were OA by virtue of their authors paying for their publication. Given that the articles were published after peer-review, any 'early view' bias was minimised. These

'gold' OA articles were available from the publisher's web site and subsequently as a matter of course, all articles become freely available to non-subscribers six months after publication. Overall, Eysenbach found, that even when taking into account and controlling for such confounding factors as the number of authors, country of origin and discipline that OA articles were still twice as likely to be cited as the non-OA articles appearing in the same journal. Eysenbach took a sample of 1492 articles published between June and December 2004; of these, 212 were OA at the time of publication. He then recorded their citation counts at three intervals over a period of sixteen months from December 2004 to October 2005. At April 2005, he found that 49% of the TA articles remained un-cited as against 36.8% for the OA articles, with this trend continuing until October 2005. Whilst these results look impressive, it should be borne in mind that the peak of citation counts generally occurs at about three years after publication (Moed 2005, p.95), a point acknowledged by Eysenbach (2006, p.696). Knowing the publication date of these articles and when they were made OA, is, however, an advantage since, for example, the studies by Harnad & Brody (2004) and Antelman (2004) do not have this information, so eliminating the possibility that any citation advantage is the result of early access is difficult to establish. Similarly, the study avoids to a certain extent the self-selection and quality bias that may be evident if OA articles are taken from repositories where authors have self-archived their own work. In order to try and control for any quality bias, Eysenbach conducted a logistic regression analysis to control, for example, for author seniority. He also conducted a survey of participating authors and asked them for their perception of the quality of their work. There appeared to be very little difference between the perceived quality of OA or TA articles. Given the high status and quality of the *Proceedings of the National Academy of Sciences*, the results found by Eysenbach are not necessarily applicable to journals containing articles of a more variable quality.

In conclusion, Eysenbach (2006, p. 697) also suggested, that "…publishing papers as OA articles on journal sites facilitates knowledge dissemination to a greater degree than self-archiving…". He also contends that this leads to an advantage in the number of citations for those articles published on journal websites as opposed to self-archived articles. This was a contentious view and was disputed by Harnad, who pointed out that the result is limited to a single journal and that this result might be due to readers accessing the journal site first rather than trying to find other sources, given the prestige that the journal

enjoys. Davis (2007) has observed that the OA articles in Eysenbach's study were featured more prominently and promoted by *Proceedings of the National Academy of Sciences* on its front cover when compared to the TA articles (3.3% vs. 1.4%) and that such coverage amplifies the chances of articles being cited. He also found it difficult to believe that any citation advantage could be attributed to those articles that were OA given the high circulation of the journal, its preferential rates of subscription for smaller institutions and its policy of allowing immediate access for poorer countries. The argument is supported by the fact that all articles in *Proceedings of the National Academy of Sciences* are freely available after six months, but despite the relatively level playing field created by this process, an advantage is still evident. Eysenbach (2007) himself counters this argument by suggesting that OA facilitates media coverage rather than the other way round, and that the difference between the rates of front cover promotion is statistically not significant.

4.17. Research impact across multiple disciplines

Hajjem, Harnad and Gingras (2005) undertook a ten year comparison based on the work of Harnad & Brody (2004), but this time looking at 10 disciplines. A similar effect was found, but the range was noticeably smaller, giving a citation advantage within the range of 0.25 – 2.5, with law at the low end and biology at the high end. Some of the criticisms of the earlier work have been dealt with here; the number of un-cited works is given for both OA and non-OA with a small percentage (4%) of the OA articles receiving 16+ citations each. The authors also detected a steady rise in the ratio between OA and non-OA articles in favour of OA. As in the earlier study by Harnad and Brody (2004), an electronic robot driven by a computer algorithm was used to trawl the databases. The use of this robot was carefully justified with its programming explained and its accuracy quantified. Nevertheless, this process has been criticised by Goodman, Antelman and Bakkalbasi (2005) who, with the agreement of Harnad and his colleagues, tested a sample from the results obtained by them. Taking a 1% sample from the records relating to biology and 8% from sociology, they found very significant miscoding of articles that were initially identified as either OA or non-OA or were missed altogether. Such was the level of error that the results obtained by Harnad and his colleagues look to be seriously flawed. Goodman *et al.* conclude that it is only possible to obtain the necessary accuracy:

- with manual determinations (which are too tedious for practical use) or

- with well-defined searches in well defined fields (such as particular journals or repositories).

The above results from Goodman *et al.* have been vigorously debated on several discussion lists. Harnad (2006) has, however, through a number of postings defended the performance of the algorithm on the basis that it is not the absolute accuracy of the robot that is at issue, even though he concedes that some errors are evident. Rather, it is more important that a demonstrable OA advantage has been detected and that this can be used as an incentive to promote OA. From those involved in the discussion, it is clear that the findings of Harnad *et al*. using the robot have been undermined. In later work, Hajjem & Harnad (2007) retested the robot's accuracy and concluded that the sample tested by Goodman was not taken from the same population as that sampled by the robot, hence the two samples were incompatible. The work by Hajjem & Harnad suggested that the robot's accuracy was sufficient to test the OA status of an article.

Antelman (2004, pp.372-382) has also found a significantly increased research impact for those articles which are freely available compared with those that are not. Taking ten high impact journals from each of the following subjects: philosophy, political science, electronic engineering and mathematics, she analysed articles from each to see if the citation counts were higher for those articles that were freely accessible online. Citation counts were taken for each article from the ISI *Web of Science*, self-citations, and citations from within the same journal and from 2004 were excluded. A search on *Google* using the article title as a phrase was used to establish whether there was a freely available online version of the article. A comparison was then made of the citation counts to see whether there was a citation advantage for those articles that were freely available. The advantage for freely available articles ranged from 45% in philosophy to 91% in mathematics. From a sample of 50 of the OA articles from each subject, Antelman (2004, p.376) found that, with the exception of mathematics, most of the online versions existed at an author's website. For mathematics, 30 of the online versions were found within a disciplinary repository. An interesting relationship between this fact and its higher impact is apparent. Antelman's methodology is described in some detail and relies, unlike the studies of Harnad and his colleagues, on the manual collection and examination of OA and non-OA records. Adding to the discussion of what motivates an author to self-archive their work, Antelman (2004, p378) noted that "the typical practice of each individual is to post either all or none of his or her articles".

4.18. Causation

Harnad (2005g) has suggested that there may be six component factors, which could be the cause of any OA advantage. The OA advantage is potentially made up of all or some of the six components listed below.

AA: arXiv Advantage, the special advantage of self-archiving specifically in ArXiv for physicists, because it is a central point of call.

CA: Competitive Advantage, for self-archived papers over non-self-archived ones.

EA: Early Advantage (View), which potentially starts at the preprint stage where OA preprints can start accruing citations before journal publication, as in the case of articles deposited in the arXiv repository.

QB: Quality Bias: this possibly arises from authors self-archiving their best papers, which may be cited more often than perhaps lesser quality work

QA: Quality Advantage (Differential) allows high-quality articles to compete on a level playing field rather than biases occurring where such articles are behind toll access barriers.

UA: Usage Advantage, OA articles are downloaded and read three times as much.

It is the EA, QB and QA components which are most often considered to make the major contributions to any citation advantage. Schwarz and Kennicut (2004) from an examination of articles published during 1999 and 2002 in *The Astrophysical Journal* and those of which were posted to arXiv as preprints were cautious about the causes of the citation advantage evident for those arXiv posted papers. They showed that papers posted to arXiv were likely to be cited twice as frequently as those that were not, but in their analysis took note of the fact that more citations may accrue to articles posted as preprints in advance of publication. They thought that the nature of the papers posted affected citation counts, for example, those:

> …authors with new results they believe to be of special significance are much more likely to post their results on astro-ph [arXiv]. The same is true for papers with particular time critical value. These effects will always cause pre-posted papers to be more highly cited on average, and without

> independent means to rank the paper quality it is impossible to disentangle them from the effects of increased visibility afforded by astro-ph. Schwarz and Kennicut (2004)

In a recent review of the literature, Craig *et al*. (2007) examined the relationship between OA article status and citation counts. The authors divided the published work into two distinct fields, earlier work which reported the average citations counts of OA and TA articles and to a certain extent implied causality and those later articles which examined citation counts but took more account of "the critical dimension of temporal progression" (Craig *et al*. 2007, p.243). Grouped in the earlier work was the research by Lawrence (2001), Harnad *et al*. (2004, 2005 & 2006), Anderson *et al*. (2001) and Antelman (2004). The authors in this group essentially counted the citations to articles that had been made OA and compared these to TA articles that had appeared in same journal. Average citation counts were then computed for both sets of data and then compared by percentage and an advantage or otherwise found. The authors were unable to isolate what might be the cause(s) of this advantage. In the second part of the review, Craig *et al*. (2007) reviewed those studies, which attempted to disentangle the possible elements that made up the citation advantage evident from the earlier studies. A number of these studies are considered below.

Davis (2006, p.103-104) in a letter regarding an observation by Antelman (2004) on causation noted that the journal publisher Emerald, had in the past often republished many of its articles in different journals. He observed, from a random sample of articles that those articles that were republished were cited more frequently than those that were not, indicating he believed that article duplication results in greater citation rates. As Antelman (2006a, p.105) herself observes that it may have been that Emerald republished only good quality articles, introducing a quality bias and in many ways OA articles are themselves a form of article duplication. Later, Davis and Fromerth (2007), taking article-level data from four mathematics journals, 18.5% of which had been deposited in arXiv, were able to find a citation advantage for those articles that had been deposited compared to those that had not. Taking 2765 articles published between 1997 and 2005, the authors found that there was a mean difference of 1.1 citations per article for those articles deposited in arXiv compared to the non-arXiv articles. They discovered in their analysis that the articles found in arXiv had been published more recently, having a mean availability of approximately 3.1 years as against 4.6 years for the non-arXiv articles,

despite the arXiv articles having a higher mean citation count. A similar difference was evident for those articles, which had more than five citations; the mean availability was approximately 4.9 years for arXiv articles and 6 years for the others. The authors examined three non-exclusive postulates, (Open Access, Early View and Quality Differential) the authors could only find reasonable evidence to support a quality differential where more highly citable articles had been deposited in arXiv. Harnad (2007a), on the other hand, thought it more likely that "both QA and [Quality Bias] QB contribute to an open access advantage, and that the contribution of QA is greater than that of QB". Davis and Fromerth (2007) were unable to say whether the articles found in arXiv were available elsewhere in other repositories or on an author's website or indeed, whether the non-arXiv articles were self-archived elsewhere. In an intriguing article, Dietrich (2007) looked at the dependence of citation counts on the relative position of eprints deposited on the arXiv astro-ph server, i.e., where they appeared on the daily list of postings. The author found that those articles that appeared at the top of the daily list had greater citation counts than those that appeared lower down the list. Dietrich concluded that the effect was due in part at least to-self promotion by authors, but could not rule out that increased visibility at the top of the daily listing contributed to the citation advantage.

Metcalfe (2006) comparing the citation rates of solar physics articles made freely available in arXiv or in the Montana State University archive found a citation advantage compared to those articles that had not been deposited. He examined all 171 articles that appeared in *Solar Physics* during 2003. One article had been deposited in both arXiv and in the Montana State University archive, another seven were posted exclusively to the Montana archive and a further six were posted exclusively to arXiv; whilst the remaining 157 were un-posted. He used the Astrophysics Data System (ADS) to collect and count citations to these articles. Metcalfe (2006, p. 551) suggests that this citation advantage is due to improved visibility rather than authors selecting their better papers to archive. Metcalfe noted the results of Schwarz and Kennicutt (2004), who found that astrophysics conference papers posted to arXiv were cited twice as frequently as those that were not. Metcalfe sampled a set of conference proceedings from solar physics and found a comparable boost in citation rates for those that had been self-archived to arXiv. Metcalfe (2006, p. 551) suggests that conferences in astronomy and astrophysics are not affected by a self-selection quality bias because they are the place to publish work in progress or

details that are not significant enough by themselves to merit publication in a peer-reviewed journal, and so he concludes are of lesser quality. It is clear however, that these conference papers were selected by their authors as worthy of posting to arXiv, so some selection process had been undertaken. Unlike Moed (2007), Metcalfe does not attempt to quantify any early access effect, but uses the results of Schwarz and Kennicutt (2004) to suggest that there is little difference in the long-term citations patterns of papers posted before and after peer-review. The sample used by Metcalfe was very small, but he noted that of the two archives, the greater citation impact was evident for those papers posted to the arXiv repository, suggesting that its greater visibility to the astrophysics community was the reason for this advantage.

Moed (2007) looked to estimate the early view and quality bias effect on the citation impact of preprint articles found in the condensed matter section of arXiv. Taking a large sample from hundreds of journals of deposited and non-deposited articles, Moed found a strong early view and quality bias, but was unable to find a general open access citation advantage. This result was obtained by looking at 74,521 articles deposited in the condensed matter section of arXiv between the years 1992-2005, about 75% of these were linked to articles published in journals indexed by the *Web of Science* with a median time between arXiv deposit and publication of six months. Moed (2007) selected from these journals, 24 of which had 10 or more articles or one percent of them linked to arXiv. Citations were counted to these articles and to the articles in the same journal that were not archived to arXiv; citations per paper were plotted over a seven-year period for arXiv and non-arXiv papers on a three-month moving average. When the curve for the arXiv papers was translated to the right for six months, the two curves for the first 24 months were very similar both reaching their maximum and then declining as shown in Figure 4.1.

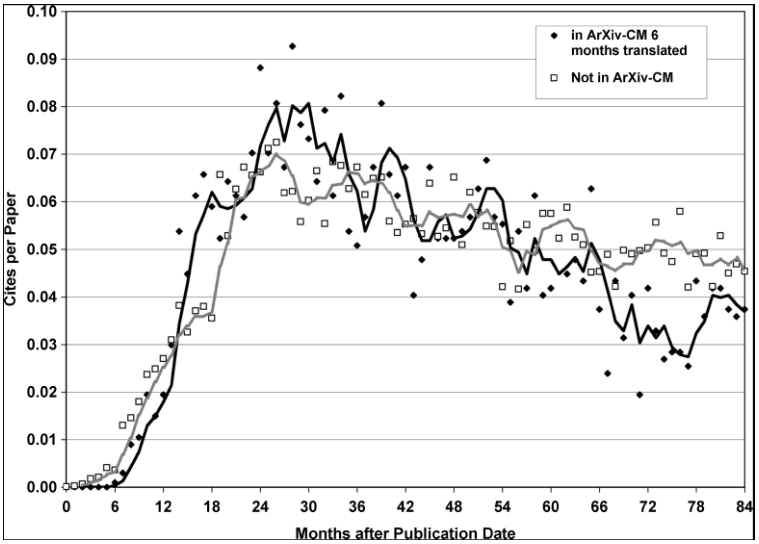


Figure 4.1 Age distribution of citations (Moed 2007)

Author prominence was evident in the frequency with which the highest cited authors had papers archived to arXiv when compared to non-deposited authors based on a single journal's publications. Author prominence was determined from the citation impact of the non-archived articles rather than from those articles that had been archived. The evidence suggests that the more senior or productive authors are more evident in the articles deposited in arXiv. After controlling for quality bias and any early view effect, (Moed 2007) suggests that there is no evidence for an open access advantage, but rather that depositing papers in arXiv accelerates citation. The findings are not dissimilar to those of Kurtz et al (2005), who were not able to discern any open access advantage in their examination of astronomy papers. It has been argued however, that the field of condensed matter is not that different from astronomy in terms of the access that scientist have to the literature, especially as like astronomy, it appears to have a fairly strong preprint culture (Harnad 2007a).

Kurtz and Henneken (2007) assert that they can demonstrate conclusively that there is no open access advantage for articles that appear in the *Astrophysical Journal*. Research that has tried to demonstrate this generally has always been hampered by some particular bias, notably for not knowing whether an author is self-archiving their better work or that better authors self-archive. The authors identified a dataset of articles from the *Astrophysical Journal* from 1997. In that year, the journal was online and fully open access. The following year, 1998, the journal erected subscription barriers and only after three years would articles become freely available. A number of authors in both years had

self-archived their work to arXiv and some had not. Taking just those authors who had not self-archived their work to arXiv provided Kurtz and Henneken with articles from 1997 which remained throughout as OA and another set of articles from 1998 which were all TA for at least three years. If there were a citation advantage to open access articles, then it would be expected that the 1997 articles would have a greater citation count. This, however, was not the case with the citation profile for both sets of non-archived articles (not e-printed) being almost exactly the same, as shown in Figure 4.2.

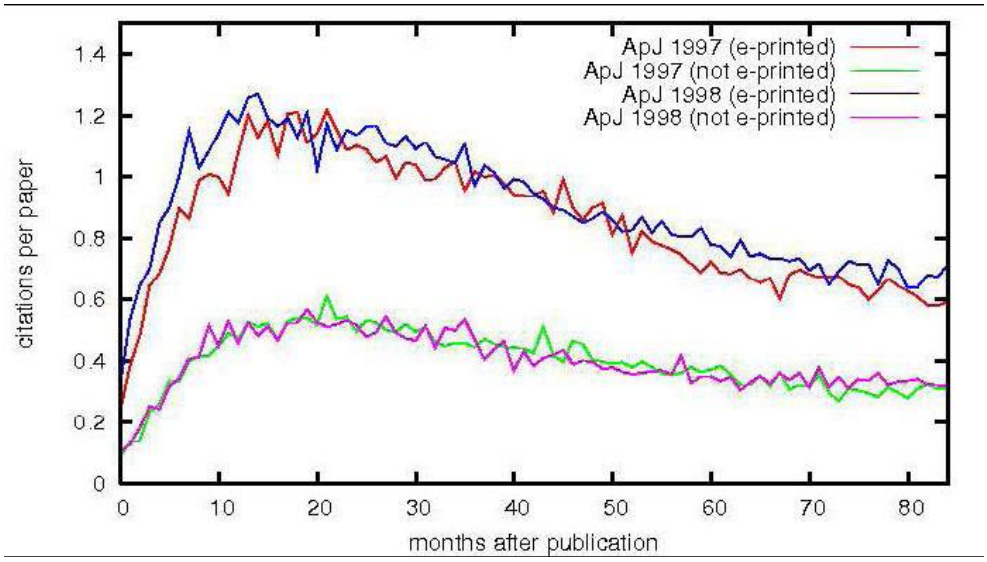


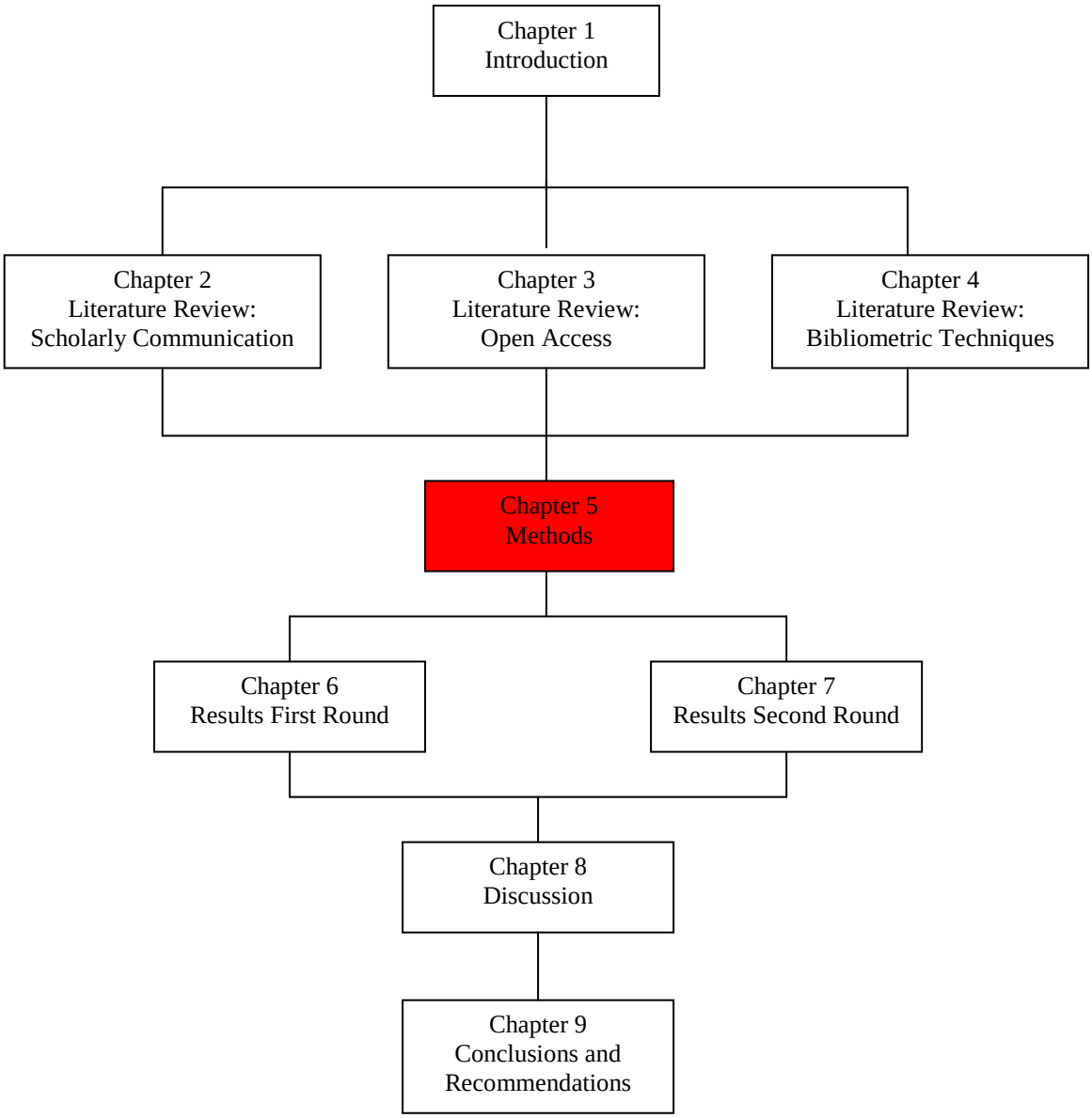
Figure 4.2 Citation profile (Kurtz & Henneken 2007)

It is also noticeable that after three years, when the 1998 articles become freely available, there is no evidence of an increase in the rate of citation. Kurtz and Henneken conclude that this demonstrates that the difference in citation rates is not due to the removal of any subscription barrier. This suggests that early access or the quality of the article is responsible for any increase in citation counts. Harnad (2007b) has noted these findings and concludes, as Kurtz *et al.* (2005) in an earlier work explained, that those who are "in a well funded field like astrophysics …who [are] in a position to write research articles has full access to the literature". Hence, Harnad (2007b) argues it must be the case that where access is not a problem, then an increase in higher citation counts can only be attributed to other factors other than an article being open access. The work by Moed (2007) in condensed matter appears to confirm this finding.

Generally, earlier work that tried to identify the source of any citation advantage was not able to discriminate between the different elements that might contribute to this

advantage. This still leaves unexplained the citation advantage that many of the relatively under-funded fields and those disciplines without a recognised repository still show. Harnad (2007b) has suggested that further work, which measures the citation advantage for mandated self-archived articles from four institutional repositories, be compared to articles from the same journals for which the author was not obliged to archive their work. Eysenbach (2007) suggests that only results from a fully randomised trial can provide a conclusive answer. Craig *et al*. (2007, pp.247-248) noted the rigour of the work by Moed (2007), particularly his imposition of a defined citation window, this being the only study of those examined that could clearly identify the date of earliest dissemination of each of the articles. They conclude that the early work looked for causal links between citation counts and OA status, but this approach while attractive, has been overtaken by methodologies that are more sophisticated. These are starting to take account of the subtleties and complexities of the different subjects examined and to "dissect the factors that drive the observed correlation" Craig *et al*. (2007, p.247). In conclusion, they recommended that the bibliometrics community design methodologically sound studies to assess any relationship between OA and citation counts across a number of varied disciplines.

Chapter 5 Methods



5.1. Introduction

To establish whether there is a citation advantage for articles that are open access (OA) as opposed to those that are toll access (TA) and make some judgement about the validity of such a claim requires the collection of extensive bibliographic records and their citation counts using suitable data collection methods. Such data, once collected, lends itself to statistical analysis. Citation analysis is one technique among the family of techniques covered by bibliometrics. This chapter briefly discusses the position of this type of research, bibliometrics, within the discipline of information science; it goes on to examine the history of the scientific method and follows with a justification of the choices made in the selection of research methods, subjects, citation database, search tools and the methods used to undertake the research and its analysis. Areas considered include the design of the research methods employed by other researchers, and the use of the literature review and pilot studies to re-frame the research questions.

5.2. Background: Information Science and bibliometrics

Information Science is generally considered to be a multidisciplinary domain sitting amongst the social sciences (Summers *et al.* 1999, pp.1156-1161). White and McCain (1998, p.334), used author co-citation analysis to map the activities of key authors within the discipline over a 23 year period (1972-1995). Citation analysis, bibliometrics and citation theory were amongst the twelve factors they identified. Other topics such as user studies, communication theories and science communication were also evident. The diagram below illustrates the place of information science as described by Summers *et al.* (1999, p.1159), who justify its place by suggesting. "If information science is to provide explanatory models, its practitioners must expect to be multidisciplinary"…[and that]…"The multidisciplinary nature of information science is clearly represented by the 'soft' disciplines that characterize the user through to the 'hard' disciplines that provide the tools"

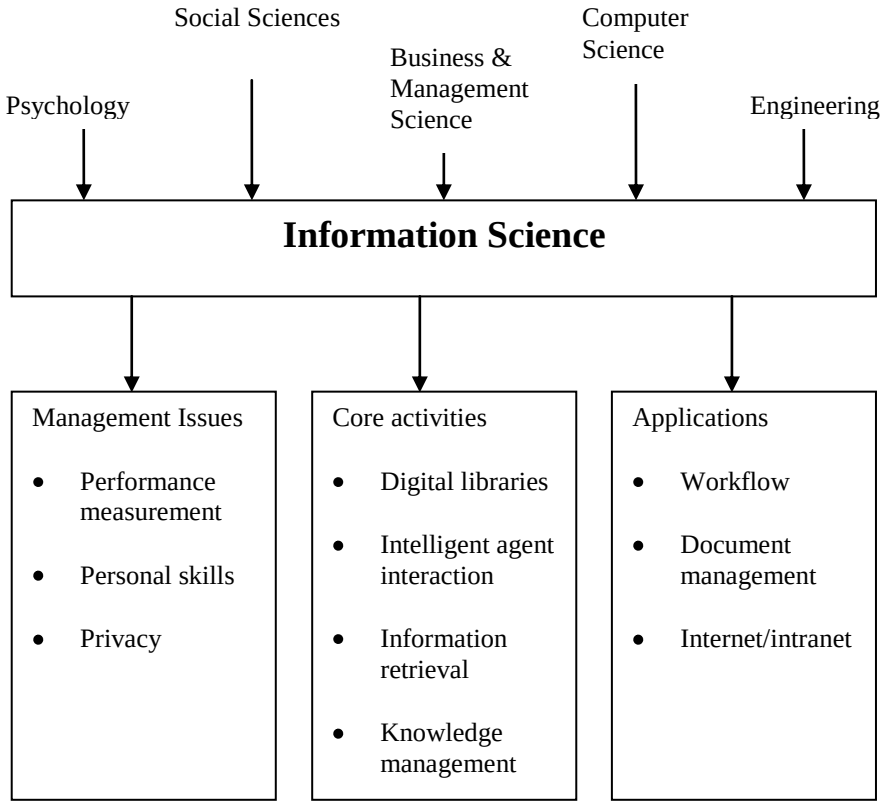


Figure 5.1 The field of Information Science

Rousseau (1994, pp.385-386) thinks of bibliometrics, scientometrics and informetrics as being sub-fields of information science and, although writing later than Pritchard below, draws the same analogy with econometrics. Here, economics is the parent discipline and econometrics the subfield which is dominated by the application of mathematical techniques, a process Rousseau (1994, p.386) thinks will "pervade the information sciences". Bibliometrics remains a key subject area and may be reasonably placed within information science, which in turn sits in the mainstream of the social sciences. However, as Borgman (1990, p.13) notes, "bibliometrics consists of empirical research methods and does not necessarily have any social science content". Whilst this is true, the actual act of citing is considered a social act and there are many reasons why people cite the work of others (Baird and Oppenheim 1994, p.6). As reported earlier, Pritchard (1969, p.348) thought of bibliometrics, like Rousseau, as analogous to 'biometrics' 'econometrics' and 'scientometrics' and would shed:

> …light on the processes of written communication and of the nature and course of development of a discipline (in so far as this is displayed through written communication) by means of counting and analysing the various facets of written communications.

Pritchard places emphasis on "counting and analysing", indicating a quantitative approach within the subject. Also Lancaster, in Diodato (1994, p.13), defines bibliometrics as "The application of various statistical analyses to study patterns of authorship, publication and literature use…". Hence, it is reasonable to accept that whilst the subjects found in Information Science locate the discipline in the social sciences, the research methods adopted in bibliometrics can to a certain extent be related to those used in the natural sciences.

5.3. Background: brief history of the scientific method

From the definitions given earlier of bibliometrics and the use of citation analysis as an analytical tool, it is apparent that in this particular field of study a quantitative deductive approach is used, rather than an inductive one. The deductive approach has the following property: when a premise, is true then the conclusions must be true (Okasha 2002, p.19). On the other hand, in the inductive approach "we move from premises about objects we have examined to conclusions about objects we haven't examined" (Okasha 2002, p.19). Given that the principal measure of OA citation advantage is the number of citations that OA articles receive, it is also logical to correlate this number of citations with the variables outlined. Whilst it appears sensible and straightforward to approach this problem in this way, it is a misconception to think this type of strategy has always been used, or that results from it can be simply considered as adequate proof of the existence of any OA advantage, if found.

Adopting a formal scientific method where hypotheses and experimental evidence are compared and conclusions drawn from that evidence is comparatively new. As Gribbin (2001, pp.68-69) remarks, until the 17th century science was based more on mystical conjecture than evidence. Gower (1997, pp.37-38) argues that Galileo was amongst the first to develop and use scientific methods in combination with deductive logic, helping to bring together "mathematical and experimental reasoning".

This does not mean, of course, that nothing like this had been written prior to Galileo; Gauch (2003, pp.160-165) gives a history of deductive logic from Aristotle onward from which much geometry, mathematics and science was derived. Francis Bacon, writing in the 17th century, however, thought that little progress had been made in science because of inadequate methods and considered that experimentation should play a much larger part in the progress of science (Gower 1997, pp.40-61). Bacon wanted, in what he called his 'Great Instauration', "to complete a total reconstruction of the sciences, arts and all human knowledge, raised upon proper foundations; this would be a six part endeavour replacing the traditional Aristotelian account of reasoning in science" (Gower 1997, p.41). This was an enterprise he did not complete but he did insist that science was both theoretical and practical and in juxtaposing the study of causes with the study of effects he, as Gower (1997, p.45) suggests, did "assign a central role in scientific method to practical experimentation investigations".

Descartes, unlike Bacon, thought that one should begin with general philosophical principles and then deduce details to meet the expected result and not rely on "uncertain observations and risky inductions" (Gauch 2003, p.60) He did, however, place more emphasis than Bacon, who favoured experimentation, on the role that mathematics should play in science (Gower 1997, pp.66-67).

Conceivably, Newton laid to rest the argument about the value of mathematics in science and also reinforced the use of deductive methods. As Gower (1997, p.79) reports, Newton "gave scope for emphasis upon the use of mathematical results, as in the Principia, and for emphasis upon experimental evidence, as in the Opticks". It is ironic, then, that when Einstein's theory of relativity was verified by observation in 1919, it should undermine some of Newton's work and that this result should so impress the young Karl Popper (Gauch 2003, p.81). Popper was later concerned with what constituted genuine theories as opposed to those that were not. In this, Popper thought that the fundamental feature of a scientific theory is that it should be falsifiable (Popper 1968, pp.32-33). Theories like those suggested by Freud and Marx were, Popper claimed, not scientific theories because they were not falsifiable, that is because the theories could be used or adapted to explain any particular action (Popper 1992, p.43). Richards (1987, p.54) suggests, "The attempt to prove theories true is futile because it is logically impossible. What is possible is to deduce falsity of the theories

from singular disconfirmatory statements". The falsification of some of Newton's theories by Einstein's theory of relativity is the classic example of Popper's view that a theory can only be viewed as credible for as long as it has not been falsified. In fact, Popper thought that anything other than drawing conclusions from deductive reasoning was inappropriate (Okasha 2002, p.23).

Popper's work has been criticised for discounting the role that inductive thinking has in the scientific process (Gower 1997, pp.207-210). Popper was an exception; other thinkers such as Keynes, Reichenbach and Carnap accepted that although more risky, inductive reasoning had a place in science (Gower 1997, pp.212-213). Kuhn (1970, pp.6-9), in contrast to Popper, suggests that scientists are routinely engaged in 'normal science' and in the process of 'puzzle solving' within a common framework of understanding, which is commonly understood and accepted by those working in the subject; something he initially called a 'paradigm'. Only when those working in the paradigm find a growing number of anomalies, which cannot not be explained by the assumptions within the subject's paradigm; do fundamental shifts of understanding take place and a new paradigm emerge. In such a 'crisis' Kuhn (1970, pp.144-146) suggests that differing paradigms then compete for the "allegiance of the scientific community" unlike the notion of falsification proposed by Popper.

These philosophical uncertainties about the nature of conclusions drawn from deductive or inductive scientific processes and the concept of falsification mean that whilst evidence may be gathered and conclusions drawn, care has to be taken that conclusions are carefully considered before they are presented as providing evidence of any particular cause or effect.

5.4. Methodology: the research strategy

Creswell (2003, pp.3-4) suggests that in any research activity, a general framework can be identified to help guide the design and structure of a research strategy. Creswell (2003, p.5) goes on to describe this general framework as having three interrelated stages; what knowledge claims are being made; what strategies might be adopted: and what data collection methods and analysis will be used in the research. The knowledge claim is conceptualised by the researcher into a suitable question or hypothesis; the researcher then may consider which approach qualitative, quantitative or mixed

methods should be used; these, in turn, are then translated into suitable data collection and analysis strategies. Behind this general approach are the philosophical traditions which underpin the basic assumptions around each stage of this process, depending on the model of research adopted by the researcher. Although distinctions can be blurred, and researchers may adopt a mixed strategy in their research, Bryman (2004, p.20) characterises the fundamental differences between the quantitative and qualitative research strategies as shown in Table 5.1 below.

Table 5.1 Qualitative and quantitative research strategies

Fundamental differences between quantitative and qualitative research strategies		
	Quantitative	Qualitative
Principal orientation to the role of theory in the relation to research	Deductive: testing of theory	Inductive; generation of theory
Epistemological orientation	Natural science model, in particular positivism	Interpretivism
Ontological orientation	Objectivism	Constructionism

The hypothesis central to this work is that there is a citation advantage when making an article OA. Secondary questions or other plausible hypotheses (de Vaus 2001, p.11-16) can arise which are concerned to identify a particular causal link, in this case between those who make their articles OA and those who do not, and any particular features of those that do. It may be for example, that the advantage might simply be a facet of the author's status, the quality of the article or where they have published it.

5.5. Methodology: understanding correlation

Trying to identify the causal links between variables is a research objective here; being able to show that a change in one or several variables causes a change in another lends validity to any conclusions drawn from the research. Moore (2006, p.75) suggests that "A change in one variable is seldom caused solely by changes in one other", and that more commonly, there are several causes acting together which result in the change and that just looking at one variable can obscure true causal relationships. De Vaus (2001, p.3) illustrates this problem by suggesting, for example, that there is a correlation between the amount of damage caused by a fire and the number in attendance to quell it. However it seems unreasonable to say that the larger the number of people in attendance, the greater will be the fire and hence the damage; rather it is more likely

that the greater the severity of the fire the greater the damage and hence the greater the number of people who will be in attendance.

Researchers may gather data and apply statistical measures of correlation which can identify the strength and direction of relationships between two or several variables (Robson 2002, pp.420-421), but as de Vaus (2001, p.4) notes, it is possible to observe and measure the correlation between two or more variables, but it is not always possible to discern causes. This can be further complicated, as confusing "causation with correlation also confuses prediction with causation and prediction with explanation" (de Vaus 2001, p.4). These difficulties then lead researchers to infer the causes of a particular correlation, since in many cases they cannot observe or isolate causal links. This is not absolute, however, as Bryman (2004, p.231) suggests: where, for example, it could be shown that voting age and behaviour are related, it is reasonable to say that the way people vote cannot influence their age and so we can say at the very least that age is the independent variable. This leads to developing research designs which attempt, as far as possible, to isolate and identify causal links and hence give validity to the conclusions drawn. De Vaus (2001, p.4) cautions that "Good prediction does not depend on causal relationships. Nor does the ability to predict accurately demonstrate anything about causality".

With these considerations in mind, any causal link between OA and TA articles and a citation advantage, if one is present, appears initially as shown below between the two variables.

However, it is possible to infer other causal relationships between the two variables.

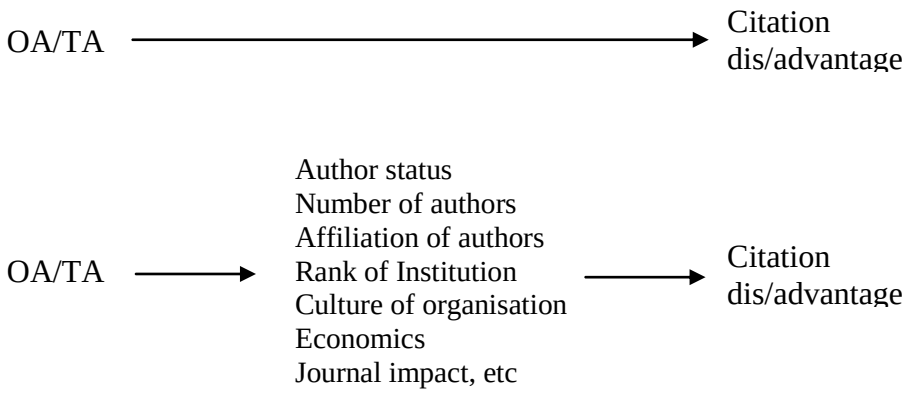


Figure 5.2 Possible causal relationships

Therefore, the hypothesis that the fact that an article is OA is the single most important factor causing an increased citation advantage is simplistic; it is possible that both this and other issues are influencing this advantage, if present. As mentioned earlier, attempting to find correlations within and between these variables is an important element in this work; clearly, however, as stated above, correlation does not always imply causality. Harnad (2005g) suggests there are at least six component factors in the OA advantage. Three of these may be particular causal factors in this advantage, **EA**: Early Advantage,**QB**: Quality Bias and **QA**: Quality Advantage

5.6. Methodology: deductive theory

Bryman (2004, p.8) considers that a deductive theory is one where the researcher "deduces a hypothesis (or hypotheses) that must then be subjected to empirical scrutiny. [and] Embedded within the hypothesis will be concepts that that will need to be translated into researchable entities." Bryman (2004, pp.8-9) outlines the deductive process as comprising stages:

1. Theory
2. Hypothesis
3. Data collection
4. Findings
5. Hypothesis confirmed or rejected
6. Revision of theory

However, the process is not always as linear as these stages suggest nor, for example, is the 'theory' necessarily something complex or large; it may simply be a generalised set

of assumptions about a particular topic. Philips and Pugh (2000, pp.16-17) amongst others are critical about the 'scientific method', suggesting there is no such thing as unbiased observation.

This deductive approach, however loosely adopted, differs from the inductive approach where theory is the outcome of the research. Here, the observations and findings help form the basis of a theory. Table 4.1 above introduces the concepts of 'epistemology' and 'ontology' as necessary to help describe the differing elements of research strategies. Epistemology is concerned with "the question of what is (or should be) regarded as acceptable knowledge in a discipline" (Bryman 2004, p.11) and ontology is concerned with the study of the "assumptions that we make about the nature of reality" (Easterby-Smith, Thorpe and Lowe 2002, p.31). Within epistemology, positivism is a position that advocates the use of methods used in the natural sciences. These are typically, for example, where the observer is independent of what is being observed. The units of analysis are reduced to their simplest forms and generalisation of the findings is typically by statistical probability tests (Easterby-Smith, Thorpe and Lowe 2002, p.30). In the case of citation analysis, the data is secondary and is completely independent of the observer and has no human dimension except in the social act of citation itself.

Objectivism is an ontological position that suggests, "that social phenomena confront us as external facts that are beyond our reach or influence" (Bryman 2004, p.16). This ontological dimension of objectivism in the quantitative approach is based on the notion that the phenomena being investigated have an existence that is independent of social actors. That is, the object of study has a clear and concrete existence which is unaffected by the conditions which surround it, or at the very least these conditions are controlled and can be quantified. This approach is clearly very different from a qualitative approach, which often involves extensive initial interpretation of the data gathered within a framework where the researcher has perceptions of their own, which in turn may affect the way the data is collected, interpreted or handled.

However, as Creswell (2003, pp.11-12) and Robson (2002, pp.42-43) indicate, researchers often adopt a pragmatic approach to their research, given that the once firm lines between the different research paradigms briefly discussed above are now less clear. This approach is characterised by researchers adopting 'what works' methods as a

fair justification to meet the goals of their research. Robson (2002, p.43-44) suggests the approach is feasible, and summarises the beliefs of Reichardt and Rallis (1994, p.85), who think that current values of quantitative and qualitative researchers are compatible and include the following beliefs:

- The value-ladenness of enquiry;
- The theory-ladenness of facts;
- That reality is multiple, complex, constructed and stratified: and
- The determination of theory by fact (i.e. that any particular set of data is explicable by more than a single theory).

Bryman (2004, pp.19-20) summarises the principal elements of quantitative research:

> Thus, quantitative research can be construed as a research strategy that emphasises quantification in the collection and analysis of data that:
>
> - Entails a deductive approach to the relationship between theory and research, in which the accent is placed on the testing of theories;
> - Has incorporated the practices and norms of the natural scientific model and of positivism in particular; and
> - Embodies a view of social reality as an external, objective reality.
>
> By contrast, qualitative research can be construed as a research strategy that usually emphasizes words rather than quantification in the collection and analysis of data.

On Bryman's analysis, bibliometrics in general and citation analysis in particular sits reasonably comfortably within the quantitative research strategy.

The quantitative deductive paradigm outlined above was used as the basic research strategy, but with the freedoms inherent in the pragmatic approach to adopt methods and data collection strategies which meet the aims of the research. Table 5.2 below gives the main elements of this approach.

Table 5.2 The research design used in this research.

What is the purpose of the research?	To determine whether OA articles have greater citation impact than TA articles and to search for causal links; using a quantitative approach.
What is the scope of the research?	The study considers aspects of scholarly communication through the perspective of Open Access and explores some causal links between certain facets of the scholarly process.
What is the focus of the research?	Examines the OA/TA impact from four disciplines as measured by citation counts from their written research output.
What are the units of analysis?	By discipline at the macro level and by articles and their citation count and bibliographic details at the micro level.
What is the sampling strategy?	Purposive and random sampling, significance defined by confidence intervals.
What analytical approach is used	Quantitative, descriptive and statistical.
Type of research	Basic research, quantitative data analysis.
Type of data	Quantitative secondary data, data-sets, numerical counts and document status data.
How were the data managed?	Secondary data sets extracted from online external databases, matched to other data and analysed to find statistical significant results.
Philosophy	Deductive, Positivism, with some Pragmatic approaches.
Research justification	To demonstrate whether an Open Access citation advantage is evident and if so, whether scholars should engage in making their work OA to gain this advantage.

Adapted from Hart (1998, p.49)

5.7. Methodology: research design

Research design is concerned with providing a framework for the collection and analysis of data. De Vaus (2001, p.9) says the "function of a research design is to ensure that the evidence obtained enables us to answer the initial question as unambiguously as possible" and "The purpose of research design is to improve the quality of our causal inferences" (2001, p.34). Consideration of a suitable design requires some attention to questions that relate to the reliability, validity and replication of the research (Bryman 2004, pp.28-30). In this process, regard is given to the use of literature reviews, pilot studies, and the work of earlier researchers as sources. These can yield the building blocks that will form a credible research design. Coupled with this was a careful review of the information resources and instruments from, and with, which the necessary data for the research was collected and analysed. Reliability and validity (see section 4.22) are concerned with ensuring that findings are consistent and stable, through, for example, the careful selection and use of databases or other search tools from which

research findings are derived (Becker & Bryman 2004, pp.172-184). Hence justification is offered for their choice as being reliable and valid instruments for the research.

5.8. Methodology: literature review

Hart (1998) explains the place of the literature review as showing:

> …command of the subject area and understanding of the problem: to justify the research topic, design and methodology (1998, p.1)…[and that the researcher has]…understood the main theories of the subject area and how they have been applied and developed, as well as the main criticisms that have been made of the work on the topic (1998, p.13).

Documents are drawn from scholarly sources, both print and electronic. Other sources in the form of commissioned reports, parliamentary evidence, contributions to discussion lists and weblogs are also used. The latter two provide immediate discussion and feedback on current issues. Much of the literature is background in nature, but in the case of where earlier research work is discussed, the methods used and the conclusions drawn can be used as a basis for the new research (Becker and Bryman, 2004, pp.69-70).

5.9. Methodology: pilot studies

Van Teijlingen and Hundley (2001, p.1) suggest that pilot studies can be used as forerunners of a major study, or can be used in "pre-testing or 'trying out' of a particular instrument", and that they can help in designing a better research process. Allan and Skinner (1991, pp.169-170) note their use in the design of "empirical investigations and… [that it is] …closely related to issues of research methodology". Peat, in Van Teijlingen and Hundley (2001, p.2), generally warns that data collected from a pilot study should not be used to verify a hypothesis or be included in the main data collection exercise. This is especially true if the results of the pilot study cause changes to be made in the design of the research. Similarly, exceptions to this rule may apply where the pilot studies yield results which are in basic agreement with later results.

5.10. Methodology: previous research strategies

All the studies found in the literature review that have attempted to test the first null hypothesis proposed in Chapter 1 have counted citations, either manually or by the use of computer algorithms which trawled a large number of records. Table 5.3 shows the principal studies which have shown an OA citation advantage.

Table 5.3 Work that has identified a citation advantage for OA articles.

Authors	Subject Areas of Study	Sample Population	Methods
Anderson (2001)	Paediatrics	Online supplement to an existing print journal	Manual data collection, *Google* to find OA articles *WoS* for citations
Antelman (2004)	Political science, mathematics, philosophy electronic engineering	Articles taken from high impact journals identified by Journal Citation Report	Manual data collection, *Google* to find OA articles *WoS* for citations
Brody and Harnad (2004, 2006)	Physics fields	CD of *WoS* database	Interrogation of web by computer driven algorithm *WoS* CD for citations
Davis and Fromerth 2007	Mathematics	Articles from maths journals deposited in arXiv	MathSciNet citations counted
Eysenbach (2006)	Interdisciplinary journal	Articles from one journal that were author pays - Gold OA	Interrogation of web by computer driven algorithm *WoS* for citations
Hajjem, Harnad and Gingras (2005)	Multidisciplinary, ten subjects examined	CD of *WoS* database	Interrogation of web by computer driven algorithm *WoS* CD for citations
Kurtz (2004), and Kurtz *et al.* (2004)	Physics and astrophysics	arXiv pre and postprint archive	Interrogation of arXiv database by computer driven algorithm
Kurtz 2007	Astrophysical Journal	OA and non OA articles	Subsets of deposited articles and non-deposited articles
Kurtz and Henneken 2007	Astrophysical Journal	OA and non OA articles	Subsets of deposited articles and non-deposited articles
Lawrence (2001)	Conference articles on computer science	120k articles from DBLP (Computer Science Bibliography)	Computer algorithm Citeseer for citations
Metcalfe 2006	Solar physics articles	Articles from one journal Solar Physics	Citations from ADS database
Moed 2007	Physics - condensed matter	arXiv pre and postprint archive	Mass download from arXiv *WoS* citations
Schwarz & Kennicutt 2004	Astrophysics	Astrophysics Journal 1999 & 2002 in the arXiv archive	Citations from ADS database

These studies counted citations that OA and TA articles have received over a given period and reported the results after some statistical analysis. The work described above uses two basic techniques. The first uses large-scale data sets from which article records

are drawn along with their citation counts, and where web searches are used to determine their OA/TA status. Such processes are often automated using computer driven algorithms. The second technique uses anything from issues from a single journal to a handful of journals in a particular subject area and are characterised by the manual collection of article records, citation counts and the identification of their OA/TA status. Lawrence, Kurtz, Schwarz & Kennicutt, Davis and Fromerth and Metcalfe drew their citation counts from a number of sources; the remainder used the *Web of Science* citation indexes. Both methods have drawn some criticism. The automated approach, has not demonstrated complete accuracy, and the other studies have tended to be small in scale and limited to a handful of subjects.

5.11. Justification: database and search tools selection

When using secondary data, the user is reliant on the data collector for the selection and accuracy of the records and how up to date they are. Citation databases are by their very nature continuously growing, with new citations being added on a regular basis. Any bibliographic database will contain errors. The *WoS* citation indexes are no exception. Its errors generally arise from authors incorrectly citing the works of others. Moed (2005, 173-179) lists the type of discrepancies; and he notes that many errors arise from the difficulty of identifying family names and capturing correct volume and issue numbers of journals. He (2005, p.175-176) reports an error level of 7.7% incorrectly cited documents, but many of the errors are minor in nature, e.g., nearly half of these errors relate to incorrect page numbering. Overall, however, errors are relatively small, and as Garfield in Moed (2005, pp.171-172) points out, incorrect citations to the Watson and Crick paper on the structure of on DNA could easily be retrieved "if one were doing an article-by-article citation analysis", which is the suggested method here. Borgman (1990, p.25) also thinks that reliability problems "can generally be identified and corrected by careful researchers". Whilst it is not impossible to replicate the results of any one study, it is unlikely that they would be identical even if precisely the same methods were used. A clear distinction arises then between the reliability of the method, (which is replicable), and the variation in results obtained if not carried at the same point in time.

Further complications arise in the selection of citation databases and search tools. There are a number of new citation databases which cover particular disciplines with varying levels of accuracy and coverage. Deis and Goodman (2005), Myhill (2005), Notess (2005) and Norris and Oppenheim (2007), have analysed a number of databases, including *WoS*, *Scopus* and *Google Scholar*, which cover the literature and citation counts related to the social sciences. It was shown that databases do vary in the quality, depth and coverage they provide. So, if an analysis were undertaken using a common set of journal articles to, say, count citations, as was by done Norris and Oppenheim (2007), there would be as many outcomes as there were databases used.

Moed (2005, p.113-114) describes the advantages of using the *WoS* citation indexes, not least of which is the frequency with which they have been used by other researchers. Moed was, however, writing this before either *Scopus*, a multidisciplinary database which covers similar ground to the *WoS*, or *Google Scholar* was properly established. Jacso (2005a, pp.208-214; 2008, pp.102-114) in his review of *Google Scholar*, a search engine similar to *Google*, its parent, but which returns scholarly information with citation counts and links to related material. It is clear that *Google Scholar* is not currently an adequate tool for citation counting as such, but may be useful to locate OA version of journal articles. This view of *Google Scholar* is also shared, generally, by others who have also found significant omissions in the coverage and recall from this database (Myhill 2005; Notess 2005). It is evident, however, that most reviewers feel that *Google Scholar* has the potential to become a useful source of scholarly information provided its shortcomings are addressed. Whilst these criticisms of *Google Scholar* are fair, Norris and Oppenheim (2007) found that in terms of finding links to individual articles taken from a common database of articles from the social sciences *Google Scholar* had a hit rate of 87%, compared to 88% for *WoS* and 95% for *Scopus*.

Like *Google Scholar*, *Scopus* has been subject to close scrutiny. Deis and Goodman (2005) compared *Web of Science* and *Scopus*, as has LaGuardia (2005). Deis & Goodman (2005) found them to be both excellent for their respective features and particular strengths in differing disciplines. Others (Burnham 2006: Dess 2006: Jacso 2004) have also reviewed *Scopus*. These reviewers generally acknowledge that both of these databases are primarily concerned with the sciences. The number of science-based records held by each greatly outweighs their holdings in the social sciences and the arts.

Nevertheless, both databases have large holdings of records in the social sciences and these are comprehensively indexed, although *Scopus* has only done this from 1996. When, however, Dess (2006) carried out citation searches for post 1996 records, *Scopus* had a small advantage of 1.2% in the number of citations that it found over the *Web of Science*. In a brief review of seventeen citation databases, including *Scopus*, Roth (2005, pp.1531-1536) concluded that overall, *Web of Science* remains the most comprehensive for citation searching. Jacso (2005b, pp.1539-1541) confirms that this database is the most comprehensive in terms of the number of records held and the number of records that can be searched for bibliographic links.

There are, however, crucial differences between *Scopus* and *WoS*. Whilst both databases enable the counting of citations, at the time this research was carried out only *WoS* allows cited reference searches. In this process, *WoS* counts the citations in the bibliographies of the core journal articles and other documents it indexes, irrespective of whether it indexes the citing document or not (Moed 2005, p.115-117). This is unlike other databases which are selective in how they count citations; thus, citation counts from the *WoS* are likely to be more inclusive in their counts. Even though *Scopus* has in 2006 added further functionality, it does not allow for the aggregation and counting of citations in a way appropriate to the studies already undertaken and the methods proposed here.

ISI also maintains *Journal Citation Reports* (JCR). This appears as an annual digest of the core journals *WoS* indexes, the journals are ranked by their impact factor (IF). This is a unique feature which makes it possible to select journals and their articles by the frequency of citation and a number of other metrics generated from citation counts. A useful feature is to link journal and article selection for examination in relation to the metrics contained within the JCR (Moed 2005, pp.91-99). This feature enables the position of journals in various subject rankings to be monitored on an annual basis and hence allows for some degree of analysis of whether the journal's relative position in the rankings can or cannot be correlated with some other measure of scholarly communication. On this basis and the other factors mentioned above, *WoS* and JCR were chosen as offering the necessary functionality with sufficient depth and coverage over a range of subjects, to both aid in the selection of journals and their articles and determine their citation counts.

5.12. Justification: selection of OA/TA article search tools

The majority of earlier studies looking for an OA citation advantage have searched the web for OA versions of TA journal articles. This has been done either manually or by trawling using a computer algorithm. The use of *Google Scholar* to find OA articles has not as yet been reported in this literature. Despite the above comments on the general adequacy of *Google Scholar* as a search tool and the experience quoted below in a pilot study, it and *Google* are used extensively to locate scholarly material. Carr (2006) reports that those making searches on the WWW for articles get to the Eprints repository at Southampton by using *Google* (76.05%), *Google Scholar* (15.25%) and *Yahoo* (4.93%).

In a small recall trial as part of this research, in September 2006, a hundred article records were taken from different subjects and used as a sample in three search engines, *Yahoo, Google* and *Google Scholar*. The article's title was entered, enclosed in quotation marks, to the search fields of each of the search engines. Yahoo was not as successful at finding hits as was *Google* or *Google Scholar*, nor did Yahoo find any hits in addition to those found by *Google* or *Google Scholar*. However, to a certain extent *Google* and *Google Scholar* were mutually exclusive and did return unique hits that found OA article records. The results obtained by using *Google Scholar*, in terms of the number of OA hits, were significantly better than those reported in the pilot study (Sept-December 2005) below. It is assumed that in the intervening period between the pilot study and this small recall trial that the search capabilities of *Google Scholar* have been enhanced.

As an additional resource to help locate OA articles, services which provide access to institutional repositories which host the self-archived work of academic authors were considered. There has been a significant growth in the number of institutional repositories into which authors can self-archive their research output and make it freely available. These repositories can have these records harvested by service providers such as OAIster. OAIster is a union catalogue of digital sources hosted at the University of Michigan; it was funded initially by a grant from the Andrew Mellon Foundation "to establish a broad, generic retrieval service for information about publicly available digital library resources provided by the research library community" (About OAIster 2007). Repositories make their records available to OAIster, who harvest "their

descriptive metadata (records) using OAI-PMH" (the Open Archives Initiative Protocol for Metadata Harvesting) (About OAIster 2007). This service currently harvests from over 900 repositories and contains over 15 million records, which are searchable from a single access point (About OAIster 2007).

OpenDOAR (About OpenDOAR 2007), hosted by the University of Nottingham, is another similar centralised access point to worldwide institutional repositories. It is the third part of the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) project. OpenDOAR, initially a directory of open access repositories, now offers a trial service to search the contents of the repositories that it lists (OpenDOAR 2006). Unlike OAIster, OpenDOAR does not search the repositories' metadata even if they are OAI-PMH compliant, but "relies on *Google*'s indexes, which in turn rely on repositories being suitably structured and configured for the *Google*bot web crawler" (SHERPA news 2006). Both of these service providers enable access to many repositories, including the major subject repositories such as arXiv and RePEc. On this basis, *Google*, *Google Scholar*, OAIster and the OpenDOAR service were used in combination as the search tools for finding OA versions of journal articles.

5.13. Justification: subject selection

In the past, researchers have chosen a single subject, a small group of subjects, or a broad range of subjects for analysis. Given that citations and OA status were to be determined manually in this study, then the range of subjects and the number of articles selected for examination had to be limited to a manageable size. To enable tests of statistical significance to be carried out on the data collected does, however, require a sample of sufficient size to be taken. Balancing these requirements led to four subjects, with samples of about 1100 records for each of the objectives listed below. Coverage of subjects within *WoS* varies between disciplines, with the sciences predominating. The database is, however, sufficiently broad to enable records to be collected from a range of subjects. Disciplines vary in their level of citedness and the coverage of the subject by journals as opposed to, for example in books (Nederhof 2006, pp.83-86). Moed (2005, pp.125-131) rates the coverage and ranking of subjects within *WoS* on a number of factors. Account was taken, in particular of rankings by:

- Overall coverage of the subject by journals; and

- Internationality as measured by country of origin of authors.

Four subjects were selected for examination. These were: applied maths; ecology; economics; and sociology. They were chosen on the basis that they presented a range of characteristics that would be sufficiently diverse, from which differing results might emerge. Moed (2005, pp.126-131) places these subjects in the following ranks:

Table 5.4 Subject coverage and orientation from Moed (2005, pp.129-130)

Subject	ISI % Coverage Of References To Core Journals	National orientation by author origin %*
Ecology	64	48
Applied maths	54	37
Economics	47	62
Sociology	27	72

*Note the subject appears in its discipline rather than being shown individually, for example, applied maths is included in the ISI grouping of mathematics.

The second column of the table relates to the percentages of references to articles published in ISI source journals relative to total references. The last column is an indicator of a journal's national orientation, defined as the share of the papers from the country most frequently publishing in a journal, relative to the total number of papers published in the journal (Moed 2005, p.131). Sociology emerges typically as Hicks (2004, pp.480-484) describes, as a discipline which is biased towards publishing a significant amount of material in books, leading to a low number of citations to core journals; that is authors cite significantly fewer journal articles than, say, in ecology. Additionally, both sociology and economics have relatively high national orientation, suggesting that articles published in these subjects are of more interest locally to the country rather than of international significance. Generally speaking, the 'harder' the science the more likely that scholarly communication will be through journals and that it will be more international in scope. For example, chemistry has an ISI coverage of 84% and a national orientation of 33% (Moed 2005, pp.129-131).

In summary, an appropriate research, strategy and design have been described for the study to be undertaken, together with a justification of the subjects chosen. The most

appropriate citation database and the electronic resources that will be used to identify OA versions of TA articles are identified. The next section describes the particular methods that were used to carry out the research.

5.14. Methods adopted

This section is concerned with showing how the methods and procedures identified earlier in the research design have been applied to meet the aims and objectives of the research. Issues of bibliometrics, citation impact, OA vs. TA and the general background to the OA movement were identified in the review of the literature. Pilot studies were used to identify methods and suitable tools as well as carrying out exploratory data collection. General and specific methods to be used in the research are explained in conjunction with the databases and search tools identified earlier as being appropriate for the task. The specific research objectives identified in Chapter 1 are then presented along with the particular methods employed to investigate them.

5.15. Methods adopted: literature review

A critical review of the literature was undertaken. Key articles were scanned for major themes that reflected the OA movement. Three basic strands were identified.

1. Themes identified related to the process of scholarly communication were:

- history and mechanics of the current system;
- how it is funded; and
- how it helps dissemination.

2. Themes related to the OA movement were:

- origins, aims and drivers:
- current developments:
- issues around its implementation; and
- models of OA, characteristics and funding.

3. Themes related to the measurement of OA impact

- citation impact of OA articles,
- bibliometric methods used to measure this impact; and
- the interpretation of causal factors.

A number of preliminary search terms were used to interrogate the databases associated with this subject, Table 5.5 below summarises this approach.

Table 5.5 Principal search terms and combinations

Principal term	Combinations			
Open access	Green	Gold	Advantage	Publishing
Citation	Advantage	Impact	Analysis	
Archiv*	Self	Institutional	Repository	Mandate
Bibiliometric*	Measures	Metrics	Advantage	History
Scholar*	Impact	Publishing	Peer review	Communication
Cost*	Open access	Repositories	Big deal	
Impact	Factor	Author	Journal	

From these searches, articles, reports and milestone declarations/statements were the principal items collected. Two alerting services (Zetoc and CSA Illumina) were used for key search terms appearing in article titles and also to view the table of contents of the principal journals. Two discussion lists were joined (Sigmetrics and the American Scientist Open Access Forum) as was a news forum (SPARC Open Access Forum) to stay abreast of the latest comment and developments in the field.

5.16. Methods adopted: pilot studies

Three pilot studies (see Appendix A for full pilot study results) were undertaken to test:

- Preliminary data collection methodologies, data sources, and data collection instruments;
- Whether there was an OA advantage to a sample of articles that were OA, compared to a similar sample of TA articles;
- Whether other causal links other than a simple OA/TA advantage is evident; and
- The time taken to undertake these tests for research planning.

Study 1 A comparative study similar to that by Antelman (2004, pp.372-382) was carried out in September 2005.

Study 2 A study into a single discipline assessing whether prolific authors were significantly more in evidence than lesser authors in terms of whether or not they had self-archived their work. This work was carried out in November 2005 and January 2006.

Study 3 A comparative study of the citation profile of embargoed journals and entirely closed access journals was carried out in December 2005

In Study 1, two subject areas were selected from *Journal Citation Reports*, i.e., six journal titles from sociology and three from neurology. Citation counts were collected from *WoS* for a number of articles from each journal and their OA/TA status determined by searching *Google* and *Google Scholar*. Results for sociology showed a small citation advantage for OA articles, whereas in neurology, the reverse was true. In Study 2, the assertion by Swan and Brown (2005, p.34) that the more prolific authors were, the more likely they were to self-archive was tested by analysing 601 articles on dyslexia for author productivity and seeing whether they had self-archived more or less than others. The evidence suggested that the reverse of Swan and Brown's hypothesis was true. The third pilot study addressed the hypothesis that when a toll access journal becomes OA after an embargo period, its articles will receive more citations than a toll access journal's articles that are entirely closed. Although there was a small difference, it was too small to be statistically significant.

The results from all three studies indicated results contrary to those found in the literature, and hence suggested that further research was needed.

5.17. Methods adopted: data collection

The aims, objectives and hypothesis for this study are given in Chapter 1. The principal objective of the research was to measure the citation advantage or otherwise of OA articles over TA articles and if there is an advantage to identify, if possible, the causes

of this advantage. In this section, the data collection strategies are given to meet these objectives.

5.18. Methods adopted: general data collection

The electronic interrogation of large data sets and the use of computer algorithms to find OA versions of articles on the web were rejected in favour of data collected and analysed manually. The necessary competencies to carry out research using automated electronic methods were not held by the writer. *Journal Citation Reports* were used to select appropriate journal titles for examination from the four subjects identified earlier. The status of each of the journals selected for examination was checked to ensure they were TA by reference to the relevant publisher's web site, and by checking them at the Directory of Open Access Journals (2006). The bibliographic records of journal articles were taken from the *WoS* citation databases; letters, book reviews, corrections, editorials and any other non-article material were filtered out. Any articles defined as reviews by *WoS* were included. *WoS* defines review articles as having more than 100 references in their bibliography.

The bibliographic record of each item selected comprised the journal issue details, the article title, authors, their affiliation and publisher. The data was downloaded and transferred to an Excel spreadsheet for initial cleaning and verification.

Spreadsheets were developed to allow the citation count for each individual article to be broken down into journal self-citations, author self-citations and others. A basic error trapping routine was built into the spreadsheets to eliminate gross input errors and negative number counts. All non-active fields in the spreadsheets were protected to prevent inadvertent data entry into incorrect fields. Hard copy forms were designed and used to collect the citation data manually. Citations for each article were collected from *WoS*'s Cited Reference Search. This covers all citations that appear in the bibliographies of the core journals that *WoS* indexes and includes in some cases citations that have questionable bibliographic data. Some judgement was therefore necessary when using the Cited Reference Search when counting these questionable citations. Using the *WoS*, each individual citation record for all of the articles from each of the four subjects was broken down into the categories given, recorded on the form and subsequently entered to the spreadsheet. The citation counts and the search for OA

versions for all of the subject's articles were carried out in discrete blocks of records such that both were completed in parallel within a 24 hour period. The results from these searches and citation checks were recorded to the spreadsheet. The entire data collection period lasted from September 2006 to April 2007.

Complete OA versions of articles were found by entering the article's title enclosed in quotation marks to the search fields of each of the search engines and the repositories. The search sequence started with OAIster, and proceeded through OpenDOAR, *Google Scholar*, and finally *Google*. OAIster and OpenDOAR were always searched, and if a result was found in *Google Scholar*, then *Google* was not searched; however if *Google Scholar* did not yield a result, then *Google* was also interrogated.

As an example, one of the journals selected, *Ecology*, for which Loughborough University has a subscription, is TA. Advice was taken from staff members of the University library and the company hosting the journal after it became apparent that access to it was possible through a *Google Scholar* search from a University terminal. The consensus view was that the computer hosting *Ecology* recognised the IP address of the University terminal and allowed access. To verify that this was the case, a PC terminal was used at a public library to try and gain access to the journal; no such access was possible, and therefore all hits from the University terminal which had the recognisable *Ecology* journal web address were discounted as valid OA hits. Where doubts existed for any other journal or other suspect web site, these too were checked and discounted if access was not possible through an external terminal. When all the necessary data was entered into a spreadsheet, it was exported to SPSS for subsequent analysis as described below.

5.19. Methods adopted: particular objectives

Objective 1: Determine the OA citation advantage or otherwise by examining the citation counts of high impact journal articles from four discrete disciplines.

Journal Citation Reports was used to identify high impact journals in the four subjects given above (See Appendix B for a full list of journals titles and sample numbers). A deliberately purposive (Denscombe 2003, p.15) sampling approach was adopted in the

selection of journal titles, since in this objective, the aim is to measure if there is an OA citation advantage, and not to determine whether the distribution of OA articles is random or otherwise. This question is addressed in Objective 4. High impact journals have by definition many citations to their articles so there would, in all probability, be a sufficient number of citations with which to make statistical comparisons. This approach to a certain extent limits the possibilities of generalising any results found to the broader population of journal articles, because it is not a random sample (Becker & Bryman 2004, p.184). But, as Seglen (1992, p.629, 635) reports, levels of uncitedness can reach 50%, with larger variations occurring within different disciplines, and Garfield (2005, p.8) has also noted that from an analysis of 38 million articles covering the period 1900-2005, about half had not been cited at all. Thus, it is quite conceivable that an effect may not be found when it is there, or that a misleading result might follow. The year from which articles were selected was 2003 and the data collection period was from late 2006 to early 2007. Approximately 1150 articles were drawn for each subject. Moed (2005, p. 95) demonstrates that, in general, the peak in citation frequency is usually achieved by the third year after publication, but there is some variation in this dependent on the discipline.

Using *WoS*, the bibliographic details of journal articles were taken from 2003. In the case of sociology, some journal articles were taken from 2002 as the number of articles per journal issue was relatively small. It would have been possible to avoid doing this by selecting more sociology journal titles from 2003, but the ranking and impact factor of these journal titles may have made comparison to the other subjects' journals more problematic. Many sociology journals have extensive book review sections.

Objective 2: Confirm that the results found in Objective 1 are not a chance event.

Whilst the results from Objective 1 may or may not reveal an OA advantage the result is only from one year. To add validity to the work the subject which was found to be least OA from Objective 1 (sociology) had the same data collection and analysis repeated as above, but for 2004. Becker and Bryman (2004, pp.183-184) advise that additional verification add to the validity of the research.

Objective 3: Ascertain whether the OA/TA citation advantage is randomly evident in a population of journal articles and whether there is an early access advantage evident from OA articles in terms of patterns of earlier citations.

The sampling methods in Objective 1 were deliberately purposive in that they selected high impact journals knowing that their articles would have high citation counts. Results are not easily generalised from such a sampling method. Taking random samples of a population lends much greater authority to any generalisations that might be made from the results obtained. Taking random samples allows an equal chance for any member of the population to be selected for inclusion in a sample. Such arrangements, Bryman (2004, pp.90-91) argues, minimise sampling error and allows inferences to be drawn from the data. One of the subjects already examined was taken (ecology) using *Journal Citation Reports* to identify all of the TA journals in its subject category for 2003, and using the *WoS* to find qualifying journal articles, data were added to a spreadsheet. Using a random number generator (Research Randomizer, 2007) and sample size calculator (The Survey System, 2007), a random sample of articles was collected. These articles were processed for citation counts and OA status and subsequent analysis using the same methods as described above in Objective 1. However, during the collection of citations, their frequency distribution by citing year was recorded for later analysis with the prospect of identifying any early citation advantage for any OA articles found.

Objective 4: Determine if lower impact journals have a similar distribution of OA/TA articles and citation characteristics to their high impact counterparts.

Taking one of the subjects (economics) from Objective 1, a similar purposive sample was taken from a range of TA journals around the mean impact factor for that discipline. Journal articles were selected and processed for citation counts and OA status and subsequent analysis using the same methods as described above in Objective 1.

Objective 5: To determine if causal links can be found between the OA/TA status of an article and the collective bibliographic details of the records associated with that status.

Two approaches were used to see if such causal links could be found. The first was to take the bibliographic data from the article records collected and using logistic

regression try to determine if there were any strong predictors of OA status amongst them. The second approach was to take the article records from applied maths and identify those who were citing them and their country of origin. The data was then analysed to see if there were discernable trends inasmuch that low-income countries might seek out and cite OA articles in preference to TA articles.

Regression analysis is a statistical technique used to measure the predictive power of independent variables to determine the value of a dependent variable. In this case, the dependent variable is whether an article is OA or TA. This type of categorical variable is termed as dichotomous or binary, given that only two outcomes are possible. In the case of predictor variables, these are for example, the impact factor of the journal in which an article appears or the number of authors an article has. Generally, "…regression seeks to predict an outcome variable from a single predictor variable whereas multiple regression seeks to predict an outcome from several predictors" (Field 2005, p.144). Regression analysis in this case expects the variables to be continuous and related in a linear fashion. This makes the use of linear regression models unsuited where the dependant variable is dichotomous. Discriminant analysis, however, unlike multiple regression, is a statistical technique that allows the prediction of category membership. The technique does have certain limitations, and in particular, it expects within reason that the variables are normally distributed although some skewness is acceptable (Kinnear & Gray 2006, p.466). The technique, however, is considered less robust than logistic regression (Field 2005, p.607: Hair *et al* 2006, p.380) and any number of qualitative predictors such as country of origin or subject (of articles) can be included (Kinnear & Gray 2006, p.477). Moreover, whilst discriminant analysis does identify group membership it is less successful at describing the structure and contribution of the different variables to the outcome. Logistic regression, on the other hand, more readily allows the contribution of each variable in the analysis to be described in terms of its strength of contribution to the outcome, the odds of doing this and the direction in which this contribution is made, that is whether the odds increase or decrease the likelihood of a particular outcome occurring (Hair *et al* 2006, p.381). A comparable trial of the two techniques was undertaken to determine that the foregoing benefits of logistic regression was evident in the output from SPPS; this proved to be the case. On this basis, logistic regression was chosen as the most appropriate statistical technique, especially as Tabachnick and Fidell (2001, p.517; 2007, p.30) suggest that it

is particularly useful if the predictor variables are continuous, discrete or dichotomous themselves or a mixture of all three. Logistic regression also has fewer restrictions inasmuch that continuous predictor variables, typically, need not be normally distributed. The technique is frequently used in the medical sciences where an outcome may be the success or failure of a treatment or if an individual has a particular condition or not, given certain observed predictor variables (Hutcheson & Sofroniou 1999, p.113). In the process, logistic regression can pinpoint those predictor variables which contribute significantly to the outcome whilst discarding non-significant predictors and this may lead to identifying the causal factors contributing to the outcome.

Kinnear & Gray (2006, p.483) note that:

> In logistic regression, pivotal use is made of a statistic called log likelihood which is written variously as -2 log(likelihood), -2LL, or -2LogL. This statistic behaves as chi-square and has a large value when a model fits poorly, and a small value when the model fits well. The log likelihood statistic is analogous to the error sum of squares in multiple regression: the larger the value, the more the variance that remains to be accounted for.

Logistic regression instead of predicting the value of a variable as in linear regression predicts the probability that a particular case belongs to a certain category. In this process logistic regression uses the maximum likelihood procedure. Based on initial observations of whether an event has occurred or not, logistic regression calculates the probability of an outcome for a particular model by assessing the value of potential predictor variables in that model. Logistic regression starts this process by taking the observed data as a constant and calculates a baseline model against which it measures improvements to the model by the addition of predictor variables. The measure used for this initial model is the log-likelihood, which sums the probabilities associated with predicted and actual outcomes (Field 2006, pp.220-221). In logistic regression, it is therefore possible to calculate the log-likelihood for models containing different predictor variables and to see how they compare to the baseline model in which only the constant is included. The baseline models used for this objective contained the actual observed frequency of OA/TA articles obtained in the first and second round of data collection.

Generally, the larger the score of the log-likelihood statistic the less well the model fits the data as there are more unexplained observations. When this score is computed as -2 log-likelihood, it gives an approximate chi-square distribution from which the significance of a value can be calculated (Field 2005, pp.221-222).

SPSS provides several different methods that can be used to enter independent variables into the logistic regression model. The default method is ‘enter’ where all the predictor variables are entered into the model in a single block. Other methods include ‘forward stepwise’ entry where predictor variables are added singly to the model successively based on their significance. In a reversal to this procedure, the ‘backward stepwise’ entry model adds all predictor variables and then removes by successive iterations any variables that do not contribute to the model. Stepwise models offer the opportunity to assess the individual contribution of predictor variables to the -2 log-likelihood score and hence some judgement can be made of the importance of each predictor variable to the outcome. Such assessments also make it possible to find in which direction the predictor variable influences the model; this may be positive or negative. SPSS also calculates the change of odds that result from a unit change in any predictor variable, values less than one result in a decrease in the odds of an event occurring and values above one increase the odds (Field 2005, pp.240-241). Depending on the entry method used, there can be a different set of independent variables selected by SPSS during processing that can be used to predict the outcome variable. The analysis was undertaken in three parts. The first was for all the first round data collectively, the second for the individual subjects from the first round and finally the individual subjects for the records collected in the second round. The forward stepwise entry method was used for the primary analysis of the first round data; this gave some opportunity to identify causal links between predictor variables and the OA status of articles. For completeness, the block ‘enter’ and the ‘backward stepwise’ method was also used to see if a degree of agreement could be obtained between these and the ‘forward stepwise’ entry model. Individual analysis of the subjects was by the ‘forward stepwise’ entry model.

An assessment of the significance of any model derived from logistic regression can be made from the output that SPSS generates after the variables have been entered. Large reductions in the -2 log-likelihood statistic indicate that the independent variables have

contributed to predicting the outcome variable; the greater the difference the smaller the amount of unexplained data is left remaining. The statistical significance of the result can be read by examining the chi-square coefficient in the *Omnibus Tests of Model Coefficients* (significant if $p<0.05$) and the *Hosmer and Lemeshow Test* (significant if $p>0.05$) tabular output from SPSS. The direction of any predictor variable (toward or away from OA) and its odds of doing this are shown in the *Variables in the Equation* table within the 'Exp(B)' column.

An issue in logistic regression is the degree to which independent variables are dependent on each other; the strength of the correlation between variables can adversely affect the model's outcome. The term used to describe this is 'multicollinearity' where independent variables can be highly correlated (Kinnear & Colin 2006, p.466). To avoid this problem the independent variables were tested in a matrix where each variable had its linear correlation tested (using the Pearson r correlation coefficient) against every other variable.

5.20. Integrity: data management

Data was downloaded onto an Excel spreadsheet, where it was checked and cleaned, ensuring only complete journal article records were selected with full bibliographic details. Stray non-article items such as letters and corrections that had crept into the data were excluded. Once citation counts and OA status were added, the records were transferred to SPSS. The exploratory function in SPSS was initially used to explore and understand the data, and three records that appeared as extreme outliers were removed (see Appendix C). Data were coded, as necessary, as SPSS can only carry out statistical tests on data that are numerical, or have been coded as such. A single master file was created in this process, and from it, a series of copies were taken that could be manipulated without jeopardising the original data.

5.21. Integrity: data analysis

The data were initially analysed using descriptive statistics; these help describe the fundamental characteristics of the records being examined. This type of analysis gives basic arithmetic results such as measures of central tendency, the range of the data, frequency distributions, variance and standard deviation. Suitable tables, charts and

diagrams were produced for inclusion in the main text of the thesis. As was noted earlier, citation distributions are generally skewed and are non-normal in appearance. Using parametric tests that rely on the data being normally distributed are usually considered more powerful than non-parametric tests with which to infer probabilities and carry out significance testing (Vaughan 2001, p.6). Despite this lack of 'normality', the central limit theorem suggests that with large sample sizes it is acceptable to assume a normal distribution and hence parametric tests such as the *t*-test can be used to measure significance (Hinton, 2001 p.55).

The data was analysed at an individual subject level and collectively to show the necessary results. Tests of significance using a two sample independent *t*-test and the Mann-Whitney test were used on the separate subjects and on the collective data to discover if there was a statistically significant difference between the mean citation counts of OA and TA articles. Although the sample sizes used were relatively high and noting the above to ensure confidence in the result the non-parametric Mann-Whitney test was used also to confirm the result. In all cases, the results were in agreement. Field (2005, p.296) explains that the "*t*-test is used in situations where there are two experimental conditions and different participants have been used in each condition" and "we look at the differences between the overall means of the two samples and compare them to the differences we would expect to get between the means of the two populations from which the samples come". Kinnear & Gray (2006, p.199) confirm that the non-parametric Mann-Whitney test is an alternative to the independent sample *t*-test.

Tests of correlation between, for example, the number of authors that an article has and its citation count, the impact factor of a journal and the number of citations that an article received was tested using the Pearson r correlation coefficient. This test measures the strength of a relationship between two variables, that is, how closely are the two variables dependent on each other (Hinton 2001, p.264). Testing a range of these variables to see if there is a relationship between them goes some way to trying to establish if there is a causal link between them.

5.22. Integrity: replication, reliability and validity

Replication, reliability and validity, are three of the most prominent tools used in the evaluation of research. Being able to demonstrate some measure of compliance with

these lends confidence to the results obtained. The work is highly amenable to replication, provided that the data used in the work remains accessible and fixed in readable form. Actually attempting to carry out the same data collection exercise and achieve the same result at a later point in time will almost certainly not yield the same results. Clearly only allowing access to the original data collected can allow replication of the results.

Reliability, de Vaus (2001,p.30) explains, is an expectation that the same reading when used on repeated occasions would yield the same result. In terms of citation counts derived from a fixed citation database, the results would be the same.

Validity is concerned with the integrity of any conclusions that are drawn from the research (Bryman 2004, pp.28-29). In terms of whether OA articles generate a citation advantage over TA articles, this is clearly demonstrable and statistical tools will support any conclusions drawn. Drawing other conclusions about causal links between any OA advantage and factors is more problematical, but some elements of the research here tried to isolate these. A number of correlation tests and logistic regression analyses were carried out on the possible dependent variables.

5.23. Integrity: pitfalls to be avoided

Data collection needs to be consistent throughout, with citation counts and OA status identified within a 24 hour period. As a necessity, data sampling procedures need to have a transparent and recognised methodological justification with a clear justification for their particular use. Great care needs to be exercised in the analysis of the data to ensure that inferences drawn are soundly based on recognised statistical procedures, with suitable confidence intervals applied to ensure results are credible. Care needs to be taken to avoid the miscounting of OA/TA articles; this is noted below in a little more detail.

5.24. Integrity: limitations of the study

Compared to the work of Harnad and Brody, this study is comparatively small scale. Harnad and Brody's (2004) research took samples, of tens of thousands of articles over a range of subjects, and showed that there is an OA citation advantage. In Objectives 1,

2 & 4, of this thesis a deliberately purposive sample has been taken to assess whether an OA advantage is evident or not, given the general distribution of citation counts. Such non-random sampling techniques make generalising the results to the wider population problematic; however, the results from Objective 3 should go some way to addressing this issue. It is the case, however, that many of the studies undertaken so far could be criticised for this lack of random sampling.

All secondary data sources will have errors in them, however carefully the data has been collected. The *WoS* citation indexes are no exception. Errors often arise principally from author's inaccurate citing practices, as discussed earlier. It is expected that in practice errors are distributed randomly and so the effect will be evenly distributed through any given set of records. Additionally, given Borgman's (1990, p.25) and Garfield's (Garfield 1979 quoted in Moed 2005, pp.171-172) earlier comments, careful counting and scrutiny of citation records can minimise errors.

Finding OA articles using general search engines is not always easy; search results may run into many pages of hits and it may not be obvious from a cursory examination which are valid hits and which are not. Considerable effort was made to ensure that publisher's sites were excluded; these generally lead to a TA version of an article with access being allowed because the host computer recognises the IP address of subscribing institutions. The other side of this problem is not finding OA versions of articles when they are in fact available. This appears at first as a positive feature, because not finding an OA article would suggest that the OA citation advantage or otherwise is being understated. However, it can be argued that an OA article that is hard to find and remains unfound will have a lower citation count than those that are easily found, and by default will become coded as TA; if this effect applied in large numbers of cases, it would widen the citation advantage of OA articles. This is an issue for all the research that has been undertaken so far and it is argued here that searches for OA articles through two general search engines, the metadata of an international repository and a surrogate search of international repositories through *Google*'s indexes have minimised this problem. A further difficulty faced by all web searches is the consistency of web links to, in this case, OA articles. Many articles which look to have viable OA web addresses have broken links, and hence were counted on the day of interrogation as TA even though on another search they may appear as OA.

5.25. Integrity: strengths of the study

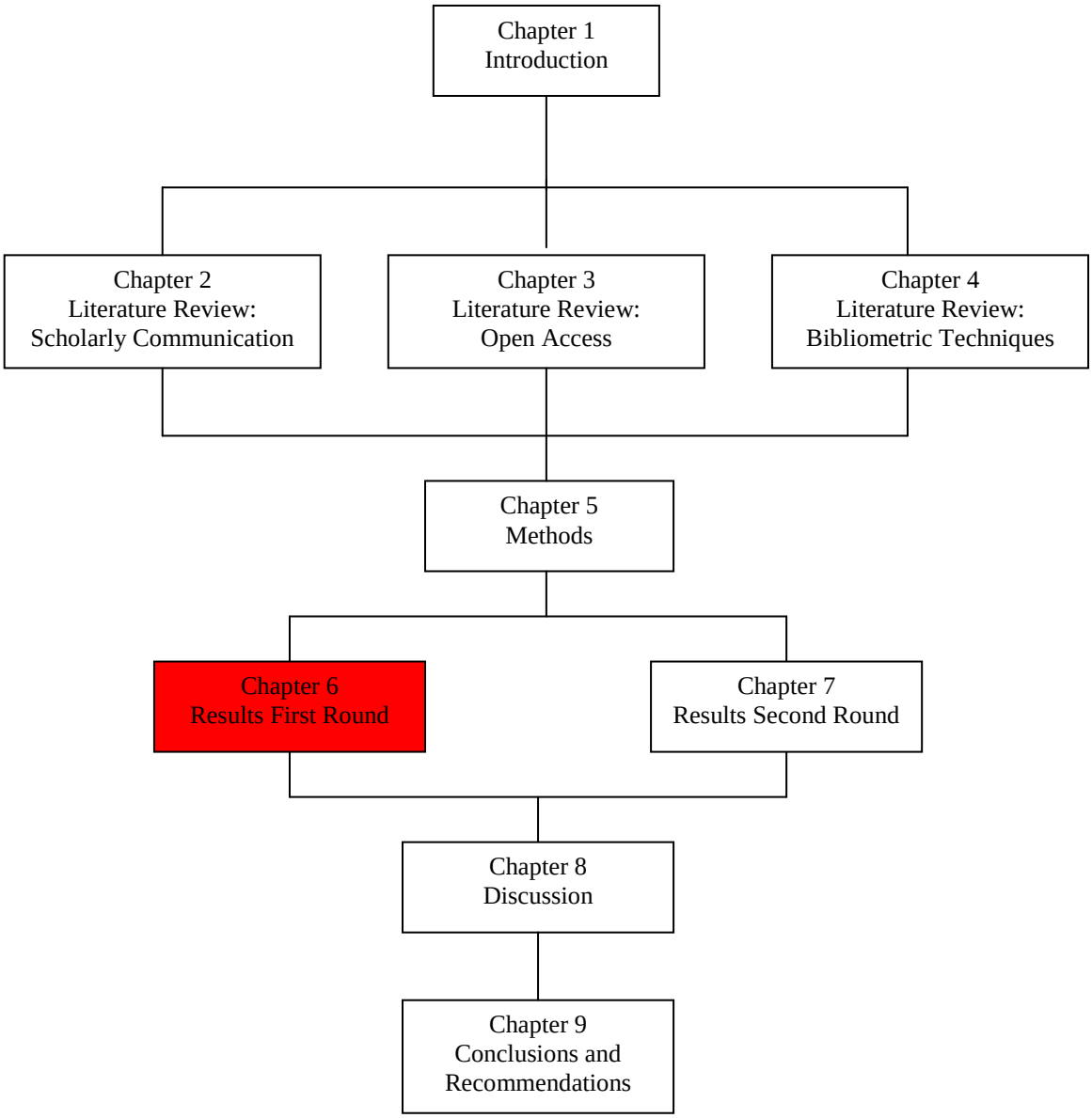
This study used the basic methods used in the majority of the other work which has attempted to show an OA citation advantage. Hence, if a similar result is evident, this would provide some further support for the results obtained to date. Whilst on a small scale, the data is manageable and the records can be examined in some detail as they are processed individually to check their OA status. However, despite its small scale, it is expected to show a statistically significant result whatever the outcome. Some of the larger studies have taken broad subject categories rather than individual subject areas, for example those of applied maths, ecology, sociology and economics that have been examined here. In taking four specific subjects, differences can be investigated in terms of their citation counts and their degree of OA, and the particular subject characteristics can be examined. This approach also allows for a tightly focussed interdisciplinary comparison between these subjects, rather than, perhaps, the larger scale studies.

Studies have tended to look at the characteristics of OA authors; here both OA and TA have been examined equally using the collective bibliographic data for each article record to identify trends. A deliberate strategy of taking high, medium and low impact journals was adopted so that differences between their OA/TA characteristics could be examined and comparisons drawn. Such an approach has not been found in the literature to date.

To date, no other study has used *Google Scholar* to check the OA status of articles; neither have the data harvesters OAIster and OpenDOAR been used extensively to find locate OA articles. Relatively little has been done to try and determine what other causal relationships might be involved in an OA citation advantage. In Objective 5, an attempt was made to try and identify some causal relationship between the general characteristics of OA articles and their authors and, if evident, any citation advantage.

The data was been collected and analysed by manual methods. Whilst this limited the volume of data collected and processed, it does allow for a visual verification of the data rather than trusting wholly to the integrity of electronic records.

Chapter 6 Results: First Round Data Collection



6.1. Introduction

Two rounds of data collection were undertaken; the first round collected citation data from articles found in high impact journals for the four chosen subjects, and this is reported in this chapter. The second round is described in Chapter 7.

6.2. Data overview

Appendix B lists the four subjects chosen; applied maths, ecology, economics and sociology, the journal titles selected from them, their impact factor and the number of articles collected from each. The total number of article records collected totalled 4633 and were split roughly equally between each of the subjects. Of the 4633 articles, 2353 were TA and the remaining 2280 were OA. In total, the articles had accrued 34,159 citations between them; 489 articles did not receive any citations at all. Three outliers were identified in the exploration of the data and they were excluded on the basis that their citation counts were too high and would excessively distort any results obtained. Their citation counts were 249, 639 and 168. All three articles were open access (see Appendix C for details). The split between the number of OA articles and TA articles varied, with sociology having the least number of OA articles. Figure 6.1 shows the proportion of OA/TA articles found by subject.

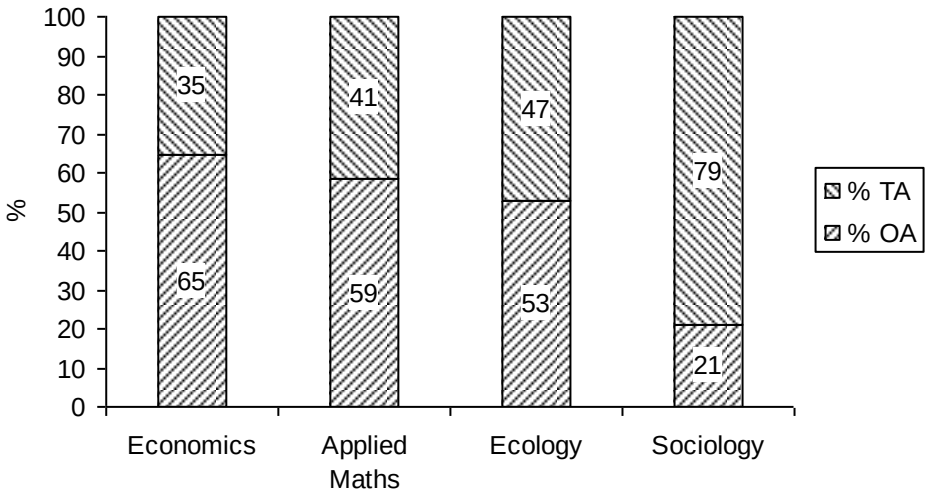


Figure 6.1 Proportion of OA/TA articles by subject

6.3. Distribution of citation counts

Figure 6.2 illustrates the positive skew of citations when all 4633 article citation records for both OA and TA records are counted and plotted. The data has a mean of 7.37 a standard deviation of 9.37 with a modal value of 1 and a median of 4. The citations counts are distributed such that 79.80% of all citations fall between 0 and 11. The overall range for citation counts is 0-117.

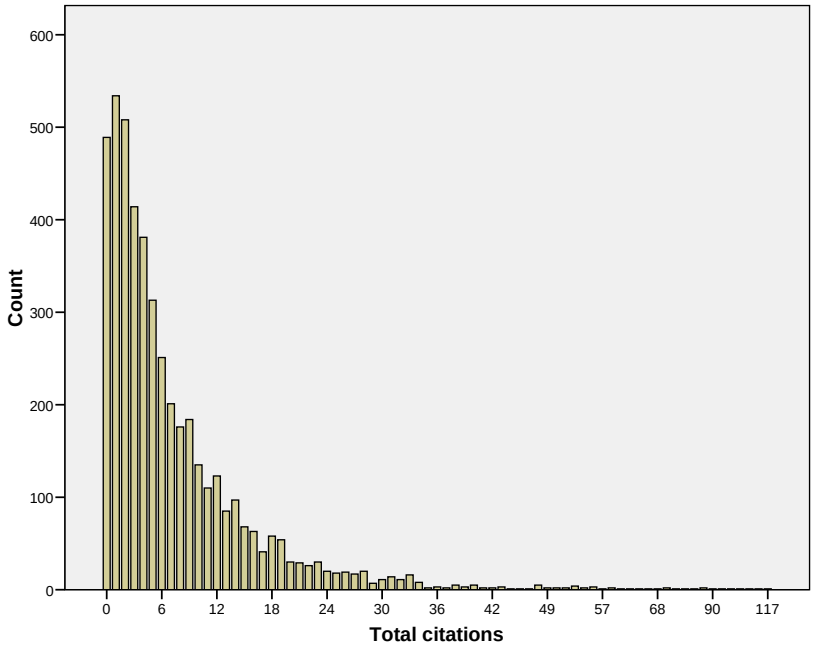


Figure 6.2 Distribution of all citations

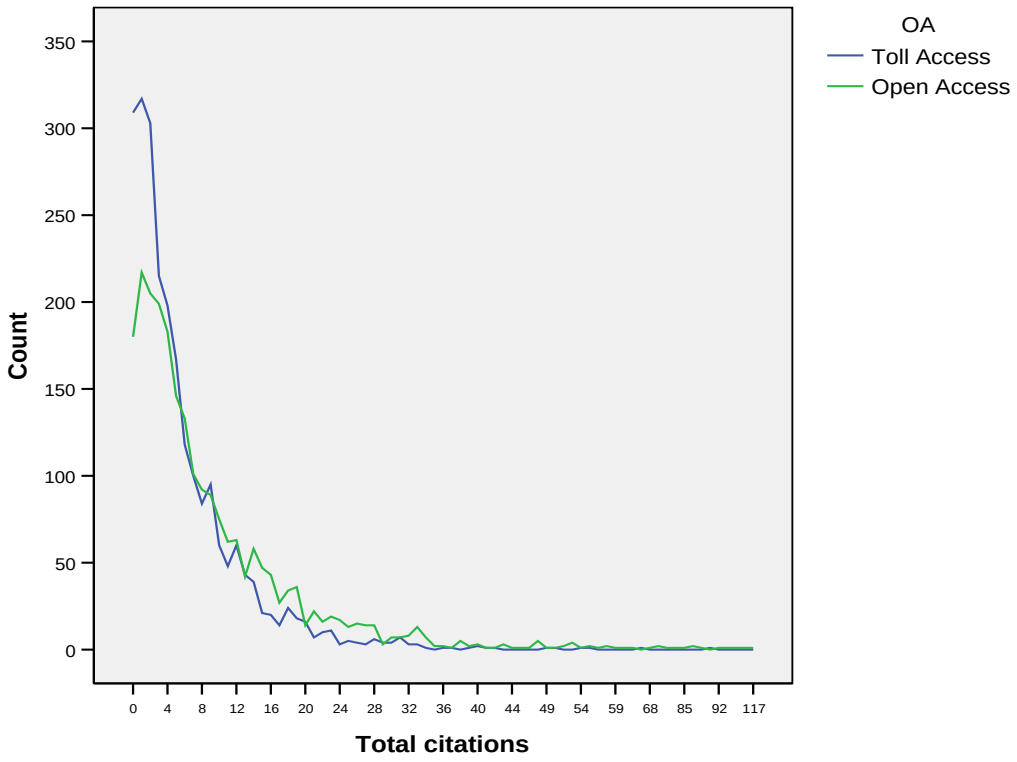


Figure 6.3 OA and TA citation distribution

The line graph at Figure 6.3 compares the citation counts for all OA and TA articles. Differences are most obvious at the zero and one-citation counts. For TA articles, 64.10% of citations fall between 0-5, whereas for OA articles, 49.60% of citations fall in this range. If individual subjects are graphed as shown in Figure 6.4 for all articles by their citation counts, differences are evident most notably for ecology.

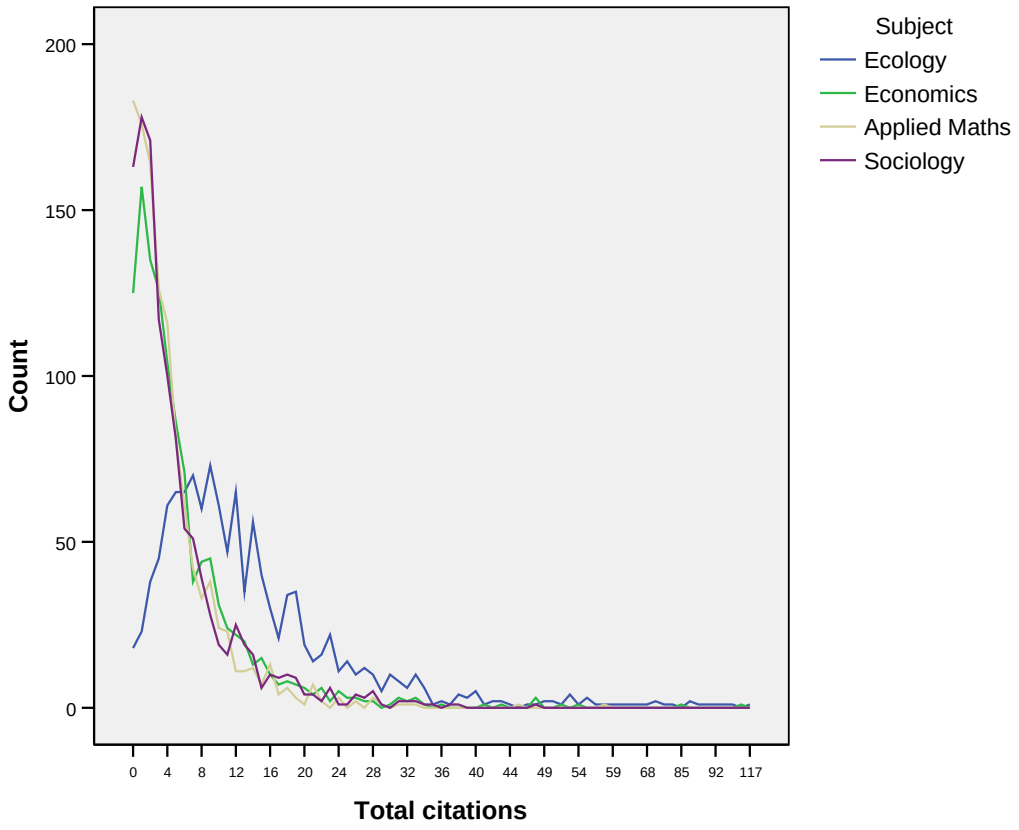


Figure 6.4 Citation distribution by subject

Of the 489 articles which did not attract any citations, the majority were TA (63.12%). Figure 6.5 shows the breakdown of these citations by OA status and subject; sociology is markedly different; it has over seven times as many uncited TA articles as it has uncited OA articles.

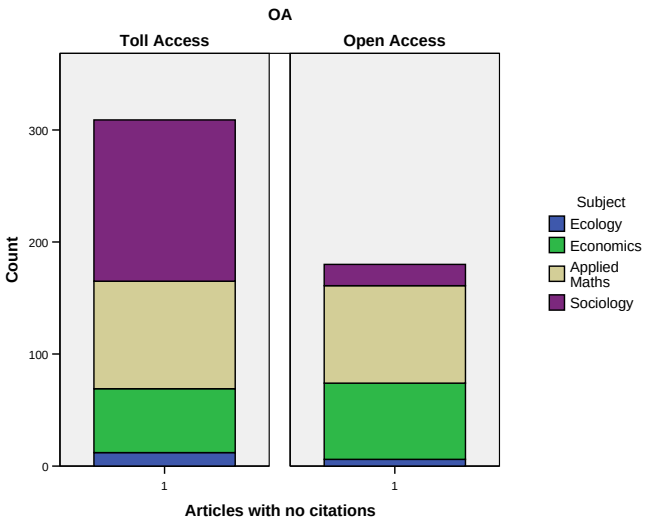


Figure 6.5 Frequency of non-cited articles by subjects

Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 9.04 compared to 5.76 for the TA articles. Table 6.1 gives the gross citation counts for the four subjects; the OA advantage ranges from 88% for sociology to 44% for ecology (OA-TA/TA citation counts *100).

Table 6.1 Gross citation counts

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	1627	480	3.39	3518	678	5.19	53
Ecology	6240	553	11.28	10012	618	16.20	44
Economics	1716	402	4.27	5099	739	6.90	62
Sociology	3961	918	4.31	1983	245	8.09	88
Total	13544	2353	5.76	20612	2280	9.04	57

The OA advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals; this is shown in Table 6.2. When these citations were excluded, the mean citation counts for the two article sets were OA 6.47 and TA 3.93. This extends the OA advantage, which becomes 103% for sociology and 49% for ecology.

Table 6.2 Citation count net of author and journal self-citations

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	854	480	1.78	2065	678	3.05	71
Ecology	4246	553	7.68	7058	618	11.42	49
Economics	1245	402	3.10	4056	739	5.49	77
Sociology	2891	918	3.15	1568	245	6.40	103
Total	9236	2353	3.93	14747	2280	6.47	65

The mean citation counts of the two populations, both gross and net of journal and author self-citation counts, were compared using the independent 2 sample *t*-test; the result showed them to be from populations with different means ($p<0.001$). Similarly when the test was conducted for each of the four subjects, the same result was found. Although the frequency distributions of citation counts are usually skewed, such distributions can, when the sample size is sufficiently large, be considered to have means normally distributed in accordance with the central limit theorem (Hinton, 2004, p. 55). The non-parametric Mann-Whitney test was also used to confirm that the two groups, OA and TA articles, were not drawn from the same populations, and in every instance, the test confirmed that this was the case.

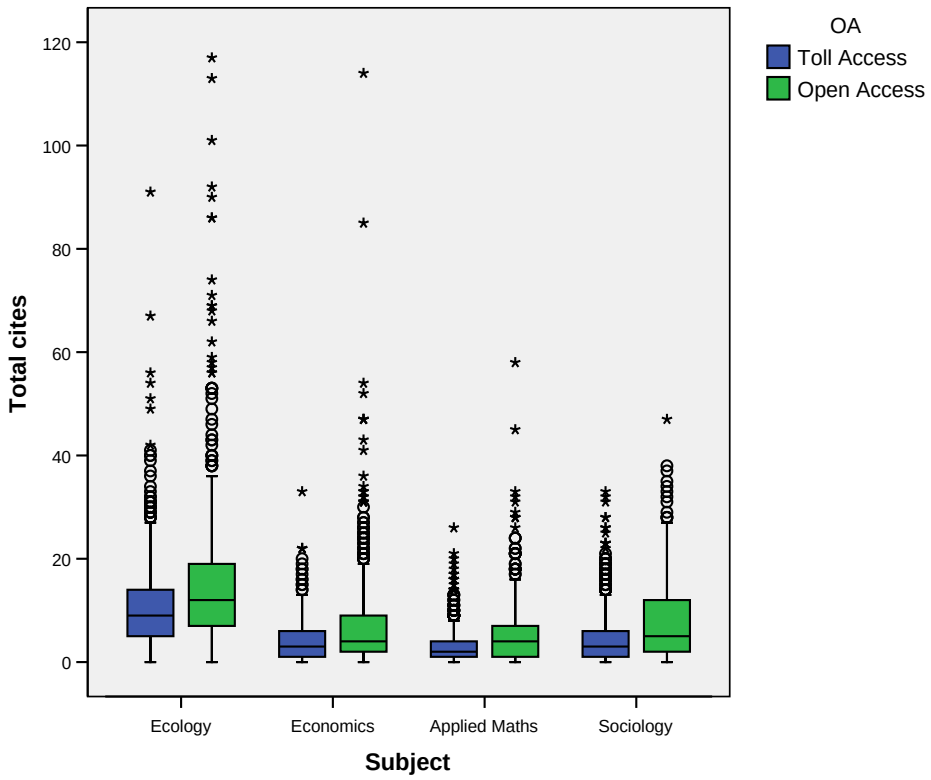


Figure 6.6 Boxplot of the distribution of all citations

The boxplot at Figure 6.6 helps illustrate the results from the *t*-test and the Mann-Whitney test showing OA articles having a greater, but variable citation median value and range than TA articles. Asterisks indicate outliers that are more than three box-lengths away from the box, whilst circles show outliers which are more than 1.5 box lengths away from the box.

For comparison, the boxplot in Figure 6.7 shows the distribution of other author citations by subject and OA status, that is only citations from other authors unrelated to the original article or citing it from the same journal. In every case, the median value for OA citations is greater as is their range.

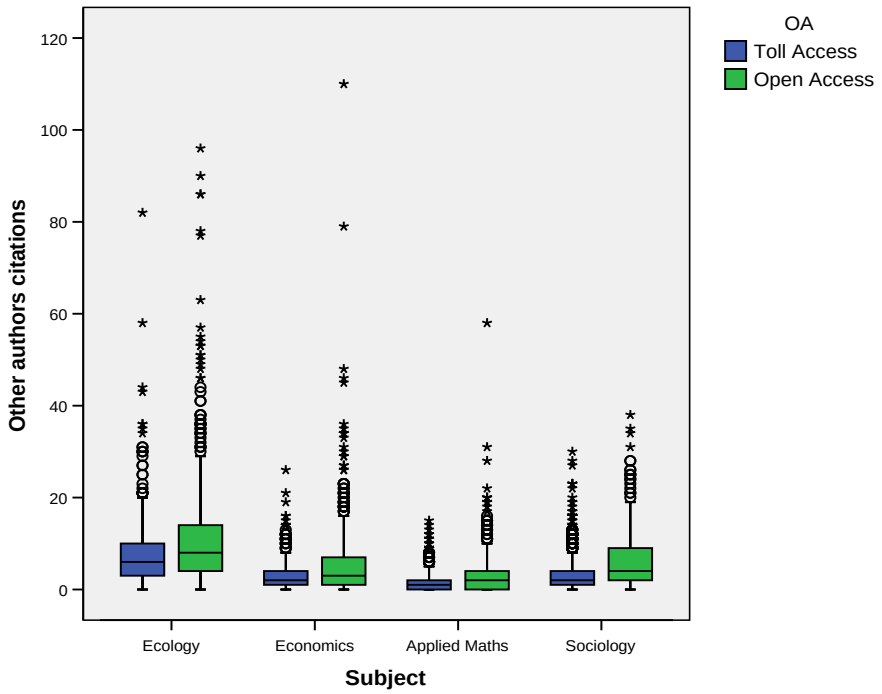


Figure 6.7 Other author citations

At subject level, the OA/TA citation distributions necessarily show some variability at the basic data level; this is shown in Table 6.3.

Table 6.3 Key citation data for the four subjects

	Toll Access					Open Access				
Subject	Mean	Mode	Std Deviation	Median	Citation Range	Mean	Mode	Std Deviation	Median	Citation Range
Applied maths	3.39	0	3.83	2	0-26	5.19	1	5.92	4	0-58
Ecology	11.28	5	9.49	9	0-91	16.20	6	14.84	12	0-117
Economics	4.27	1	4.36	3	0-33	6.90	1	9.02	4	0-114
Sociology	4.31	1	4.92	3	0-33	8.09	2	8.61	5	0-47

All categories of citation counts between 24 and 117 were combined into 10 different groups (25% of all citations) to aid a Chi-square test of association. The results from the Chi-square test ($\chi^2(32) = 194.889$, $p<.001$) showed there was a statistically significant, association overall between the frequency of citations and the OA status of an article for all categories of citations. That is, for any article with six or more citations (excluding articles with 9 and 20 citation counts) there will be a greater number of them OA than TA. The reverse is true for articles with up to five citations, including zero counts. A similar Chi-square test, which combined the citation counts between 27 and 110 into 10 different groups (15% of all citations) was performed for all other author citations. The result of the test ($\chi^2(32) = 166.887$, $p<.001$) which excluded all self-

citations showed a strong association between OA status and citation count. That is, for any article with five or more citations, there will be a greater number of them OA than TA; the reverse is true for articles with up to four citations including zero counts.

6.4. Self-citation counts

The rate of self-citation varies between OA and TA articles and within the four subjects. Figure 6.8 shows the distribution of all categories of self-citation for both OA and TA articles. The data for OA articles has a mean of 2.58 and a standard deviation of 3.25 with a modal value of 0 and a median of 2. For TA data, the articles have a mean of 1.83 and a standard deviation of 2.46 with a modal value of 0 and a median of 1.

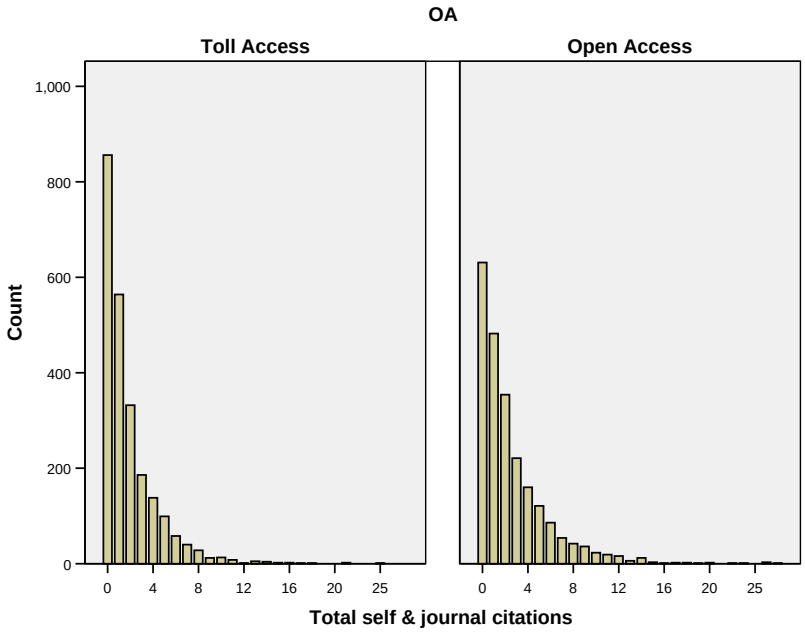


Figure 6.8 Distribution of self-citations by OA/TA status

Figure 6.9 shows the relative closeness of the self-citation counts when they are taken together. However, the most noticeable difference is apparent at the zero count, where 36.4% of TA articles have no self-citations, whereas for OA articles this is 27.7%.

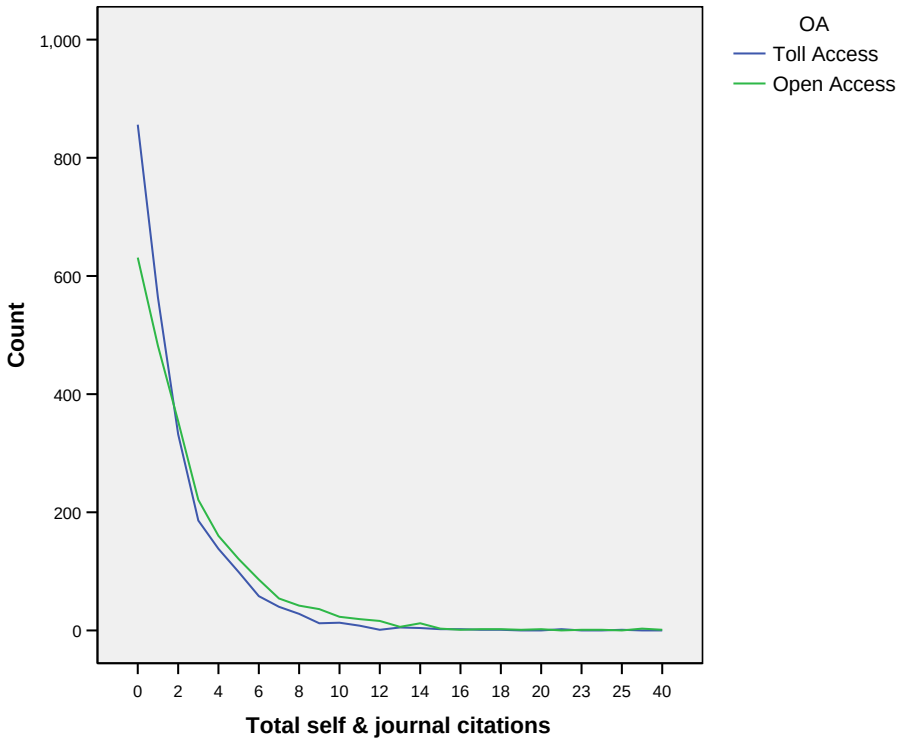


Figure 6.9 All OA and TA self citations

When subjects are examined separately by self-citation counts as in Figure 6.10, a different result emerges. There are significant differences in the frequency with which authors self-cite. In ecology, 88.3% of all the articles have one or more self or journal self-citations. In economics, this figure is 60.6%, for applied maths this is 66.7% and in sociology, 55.8% of articles had some form of self-citation. The corollary of this is that the number of articles having no self-citations is 11.7% for ecology, for economics 39.4%, for applied maths 33.3% and sociology has 44.2%.

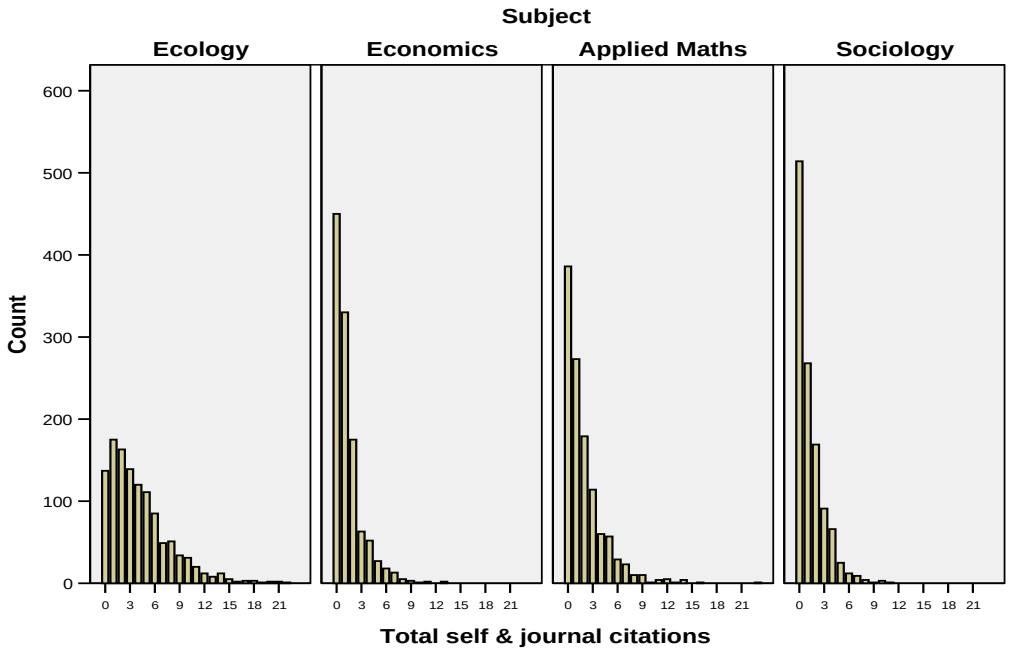


Figure 6.10 Total self-citation count by subject

The boxplot in Figure 6.11 shows the distribution of self-citations by subject and OA status. The median values for the self-citations are generally very close; ecology is the exception, having a noticeably higher median value and a greater range.

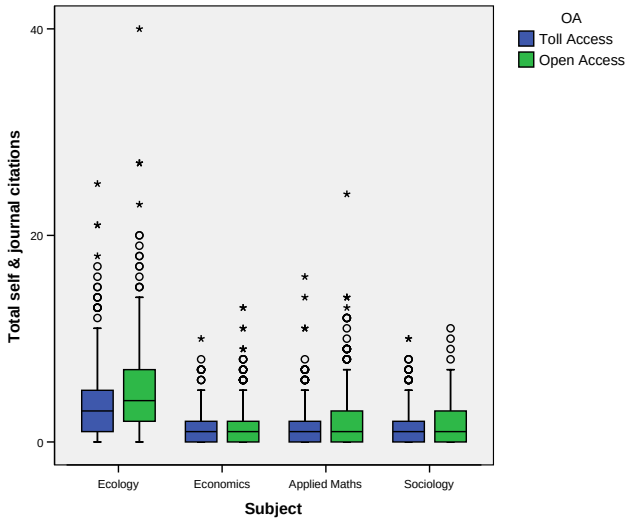


Figure 6.11 Boxplot of self-citations

Figure 6.12 and Figure 6.13 show a breakdown of the gross citation count by the four types identified, three of which are related to author or journal self-citation. Journal author self-citations (JASC) are where the cited author is citing themselves and writing in the same journal as the original cited article. Journal self-citations (JSC) are citations

where authors other than the original article author have cited the article within the same journal. Author self-citations (ASC) are where the authors are citing themselves but are writing in a journal other than the journal in which their original article appeared. Finally, other citations (OC) are from authors unrelated to the original cited journal or any of its authors.

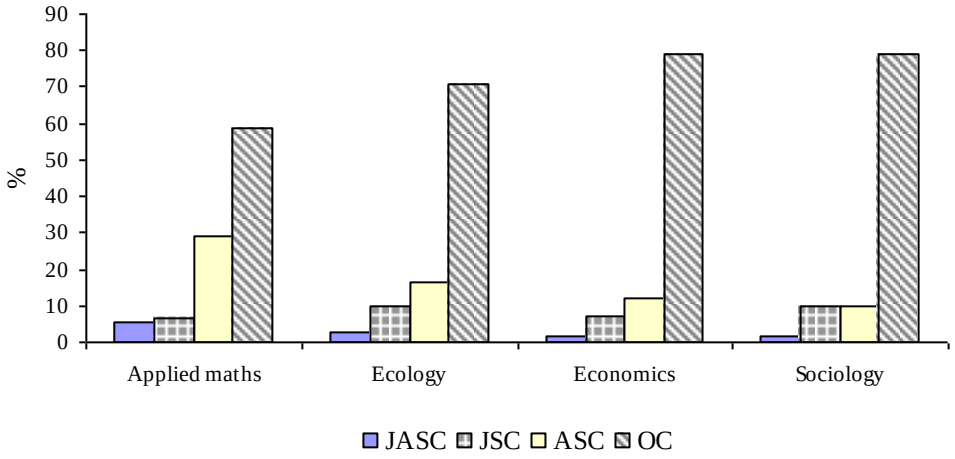


Figure 6.12 Breakdown of OA citations by subject

Figure 6.12 shows the citation breakdown for OA articles. Other citations form the largest single category for all the subjects and author self-citation rate is highest in applied maths and lowest in sociology. The combined self-citation rates for OA articles were 41% for maths, 29% ecology, and 20% for both economics and sociology. Likewise, Figure 6.13 shows a breakdown of the gross TA citation count by the four types identified. There is a very similar pattern to the citation breakdown for OA articles.

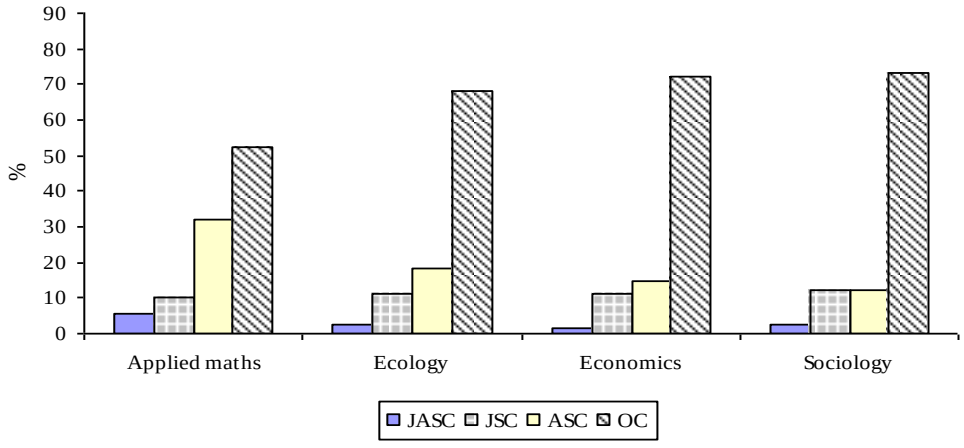


Figure 6.13 Breakdown of TA citations by subject

The combined self-citation rates for TA articles are 47% for maths, 32% ecology, and 27% for both economics and sociology. Figure 6.14 shows a comparison between each citation category and their OA/TA status and illustrates that OA articles have a greater percentage of their citations in the 'other citation' category than TA articles as a proportion of all citations within their subject. Conversely, TA articles have a greater proportion of citations in the other categories of citations.

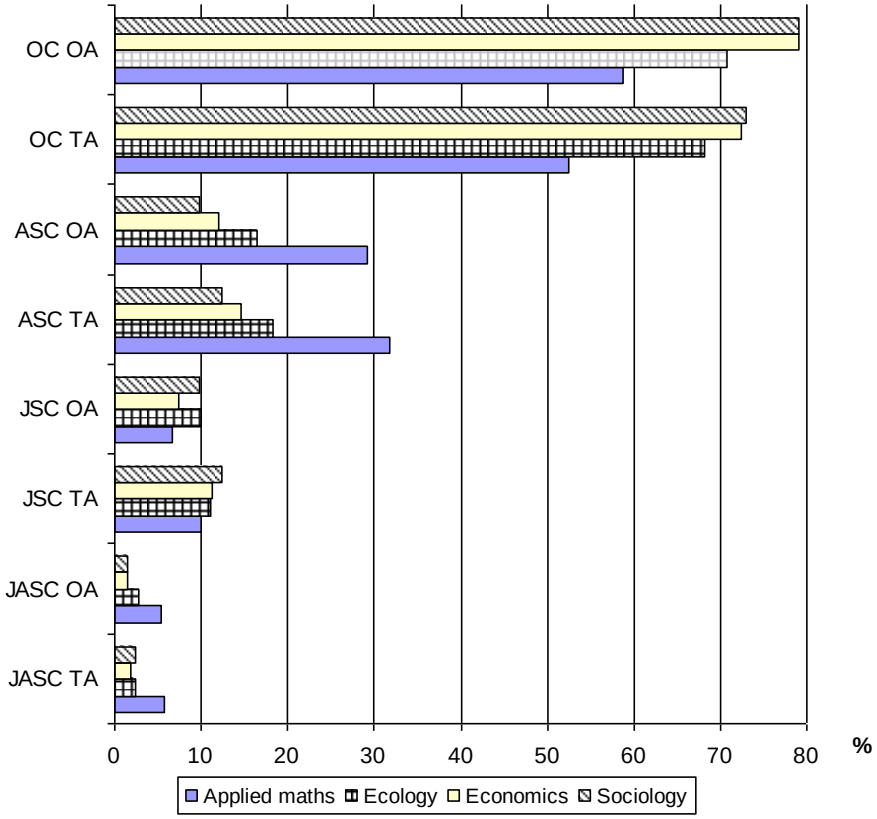


Figure 6.14 Percentage citations by their citation category

However, OA articles consistently have a higher individual citation count for the articles within the self-citation categories outlined above, despite there being fewer OA articles than TA articles in these self-citation categories. The mean number of journal and author self-citations for OA articles was 2.58, and 1.83 for TA articles. Consistent differences between the mean number of journal and author self-citations were also evident at subject level in favour of OA articles; ecology 4.78/3.61; economics 1.44/1.17; applied maths; 2.14/1.61; and sociology 1.69/1.17. These means were compared using the independent 2 sample *t*-test; the result showed all four to be from populations with different means ($p<0.001$).

The scatter plot shown in Figure 6.15 (marker size indicates frequency density) shows the relationship between other author citations (OC) and all types of self-citation. A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 6.15. The result for both was significant ($p<0.01$). For OA articles this was 0.474 and for TA articles this was 0.478.

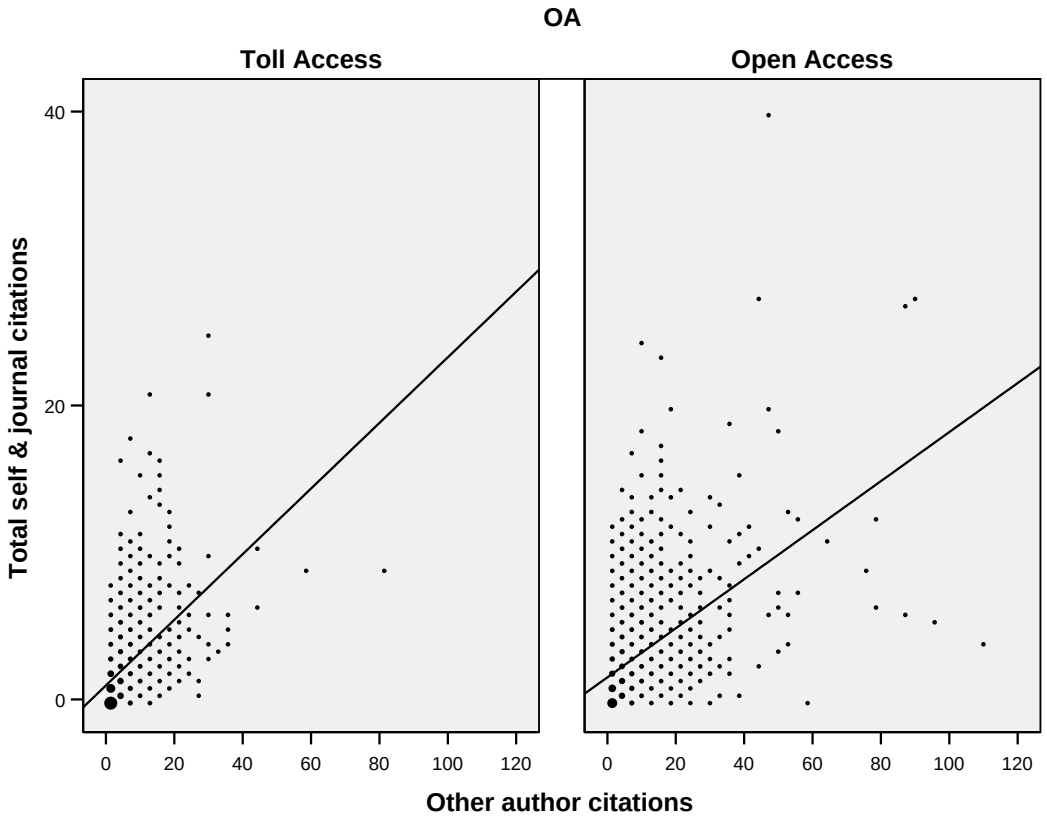


Figure 6.15 OA/TA scatterplot of self-citations to other author citations

Clearly, the rate at which other citations accrue compared to all types of self-citations varies by subject. One-tailed Pearson correlation coefficients ($p<0.01$) show that for TA articles, economics is, at 0.305 the weakest and sociology at 0.449, is the strongest. For OA articles, the weakest correlation is 0.331 for applied maths, and the strongest again is for sociology at 0.565.

6.5. Author frequency and OA/TA status

Table 6.4 shows the mean number of authors by subject and OA status. OA articles consistently have a higher mean number of authors.

Table 6.4 Mean number of authors per article

	Ecology	Economics	Applied Maths	Sociology
OA	3.25	2.02	2.43	1.9
TA	3.13	1.96	2.2	1.78

Table 6.5 shows an accumulating article count by the number of authors for each article, split by their OA/TA status. Differences are evident until the four-author level.

Table 6.5 Author counts by article frequency

Authors	Total articles	TA articles		OA articles	
		No	Cum %	No	Cum %
1	1356	832	35.4	524	23.0
2	1659	780	68.5	879	61.5
3	961	444	87.4	517	84.2
4	358	158	94.1	200	93.0
5	170	76	97.3	94	97.1
>5	129	63	100	66	100
Total	4633	2353		2280	

Figure 6.16 illustrates the distribution of author counts by the OA/TA status of their articles. There are noticeably more single authored TA (832) articles than there are OA articles (524). Apart from these single authored articles, there appears to be a leaning towards OA status for articles that have more than one author.

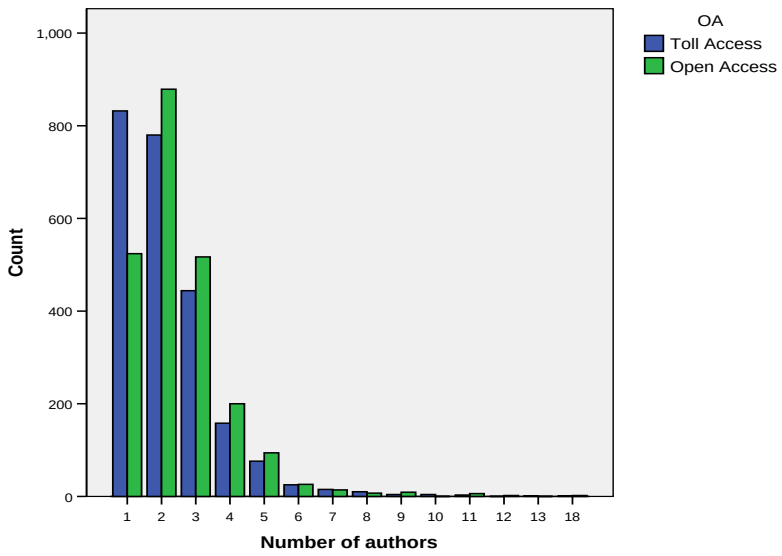


Figure 6.16 Articles by author count and OA/TA status.

Author counts of nine or more were combined into a single group (1.4% of articles) to aid a Chi-square test of association. The results from the Chi-square test ($\chi^2(8) = 88.83$, $p<.001$) showed there was a significant association overall between the number of authors and the OA/TA status of an article. However, the association between the number of authors and the OA status of an article showed that there was a tendency towards OA status only when there was more than one author. Hence, there is a strong association between single authorship and articles being TA. Of the 1356 single authored articles, 61.36% were TA and the remaining 38.64% were OA. The situation is reversed for articles having more than one author. For those OA articles having between 2-5 authors, the differences between them and the TA articles is, however, less marked, ranging, dependent on the number of authors, from 53%-56% in favour of OA articles. This result however, partially breaks down at subject level where the association is not significant for ecology or sociology, but is for applied maths and economics ($p<001$), and this association tends to be variable, favouring neither OA nor TA consistently.

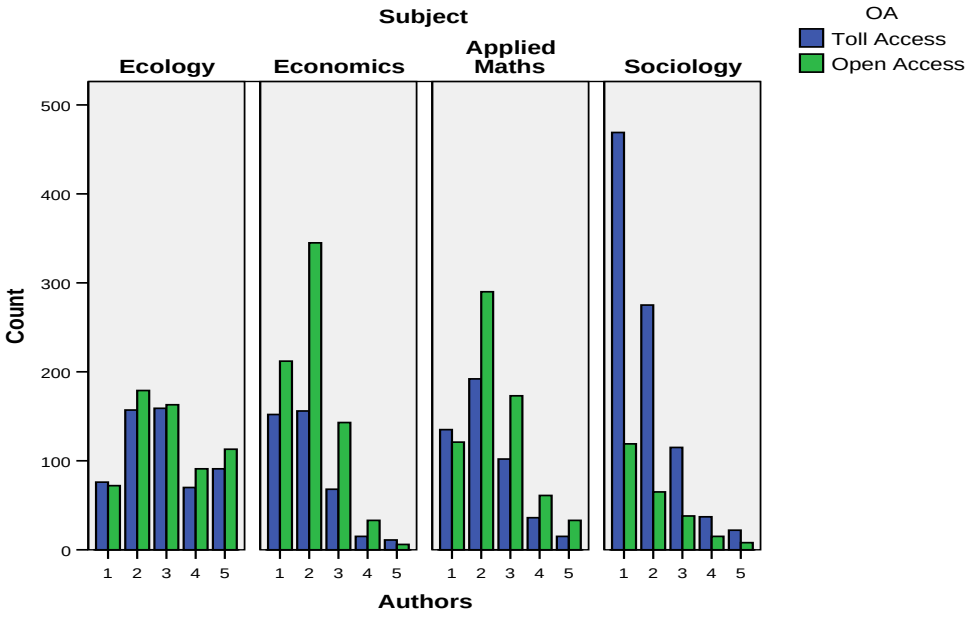


Figure 6.17 Author count by OA/TA status and subject

Figure 6.17 illustrates the variation in author count and OA/TA status by subject; articles of five or more authors 6.4% of all articles have been combined to aid clarity. Sociology shows how the distribution is heavily skewed by the predominance of TA articles and single authorship. Analysis of this data using a Chi-square test for TA articles ($\chi^2(12) = 393.99$, $p < .001$) shows a strong association between subject and the number of authors. The association is strongest for ecology and sociology. For OA articles ($\chi^2(12) = 381.93$, $p < .001$), there is a similarly strong association between subject and the number of authors, particularly at the one to two author level, with a weakening of association for three author articles but a strengthening thereafter up to the five author level.

The 4633 articles sampled yielded by first author affiliation 73 countries of origin (65 articles had no first author affiliation given). For analysis purposes, these were grouped into four regions: North America, continental Europe, UK and the Rest of the World. Figure 6.18 shows this split when author counts are limited to no more than five, six and greater account for 2.8% of articles across all four regions.

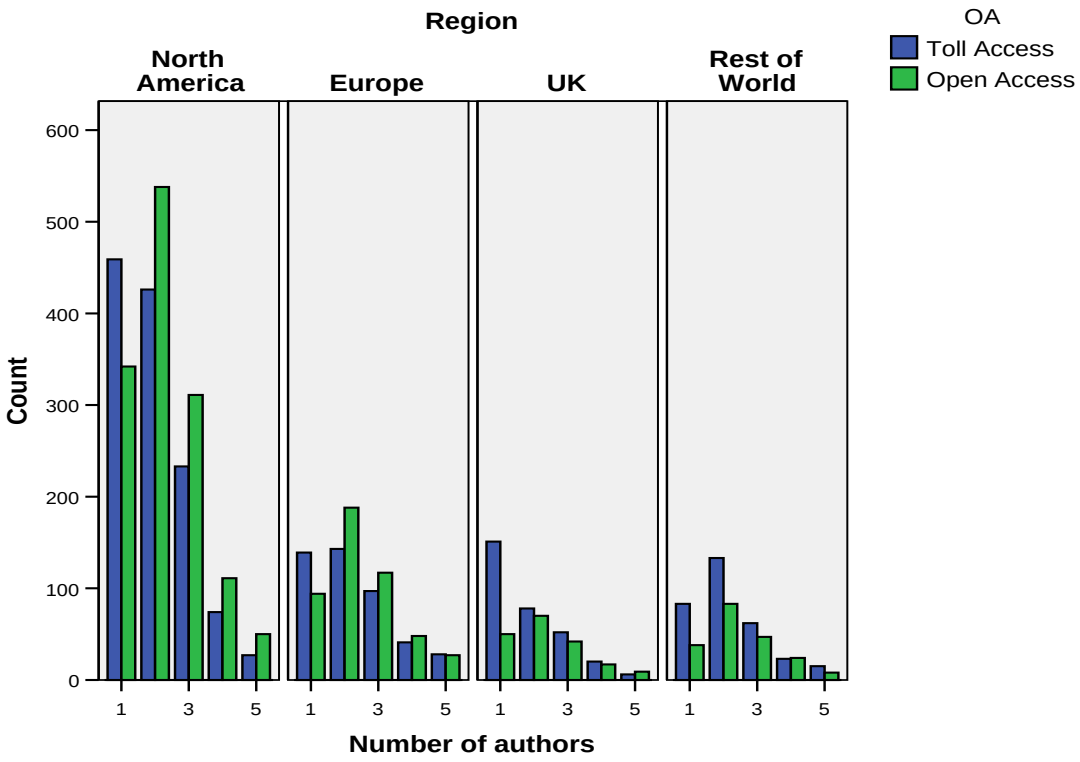


Figure 6.18 Author count by region and OA status

Taken at a regional level, the results from Chi-square tests on the number of authors and the OA/TA article status of their articles gives a similar result to that when all articles are taken together irrespective of region. For North America, there is a strong association ($\chi^2(6) = 52.24$, $p < .001$) in favour of single authors being TA and those articles having multiple authorship being OA. A similar result applies to continental Europe ($\chi^2(6) = 16.67$, $p < .001$). For the UK ($\chi^2(6) = 26.53$, $p < .001$) and the Rest of the World ($\chi^2(5) = 9.24$, $p < .001$), but the association is almost exclusively in favour of a strong association for TA articles irrespective of the number of authors. Author counts of six or greater were combined into a single group for the Rest of the World (2.27% of all articles) and for the other three regions author counts of seven or greater were combined (1.78% of articles) to aid the Chi-square tests of association.

An examination of the origin of articles by first author affiliation showed that authors from North America provided the majority of articles, accounting for 56.6% of all the articles published in the 65 journal titles. Continental Europe follows North America with 20.7% then the Rest of the World with 11.4 % and the UK with 11.3%. Figure 6.19 illustrates the dominant position of North America in the number of articles

published and the marginal advantage held by OA articles. The OA article advantage is reversed for the UK and the Rest of the World.

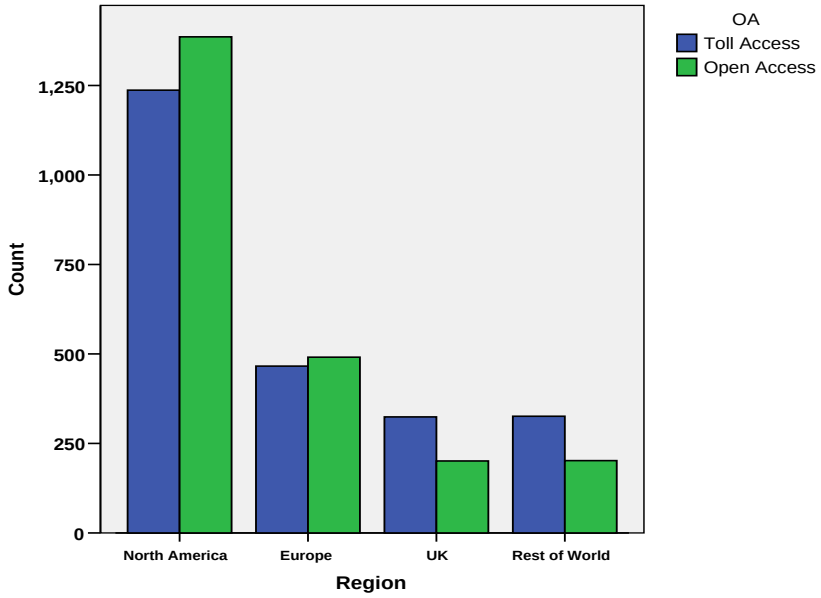


Figure 6.19 Number of OA/TA articles by region

Table 6.6 gives a breakdown of the OA status of articles by country and subject. North America has the highest rate of OA (60.8%), followed by continental Europe (21.5%), with the UK and the Rest of the World at 8.8% and 8.9% respectively.

Table 6.6 OA/TA article counts by country and subject.

		Subject and article count				
		Ecology	Economics	Applied Math	Sociology	Total
Open Access	N America	383	520	293	190	1386
	Europe	121	95	254	21	491
	UK	65	69	43	24	201
	Rest of World	49	55	88	10	202
Total		618	739	678	245	2280
Toll Access	N America	236	221	159	621	1237
	Europe	147	65	189	65	466
	UK	80	72	23	149	324
	Rest of World	90	44	109	83	326
Total		553	402	480	918	2353

The results of a Chi-square test ($\chi^2(3) = 65.922$, $p<0.001$) showed there was a significant association overall by region, subject and split between OA and TA articles. There is a tendency towards OA in North America and continental Europe, and a tendency toward TA in the UK and the Rest of the World. Figure 6.20 shows the detail of these differences by contrasting the individual subjects and their article counts by region.

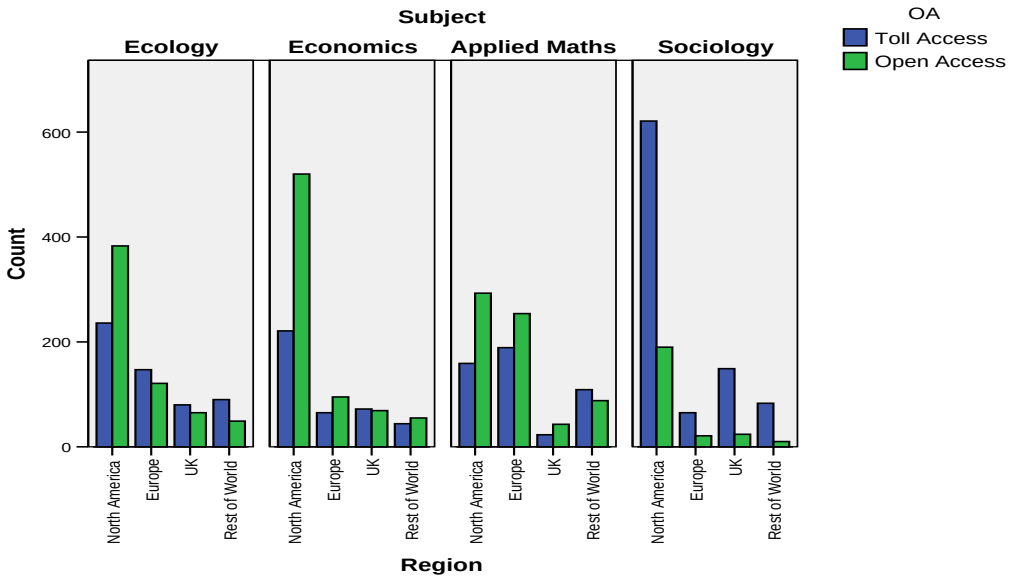


Figure 6.20 Article count and OA/TA status by region

Sociology stands out as the major exception with the majority of articles being TA even in North America, which consistently has a majority of OA articles in the other three subjects.

6.6. Correlations

The pair of scatterplots in Figure 6.21 shows the distribution of citations against the number of authors for TA and OA articles.

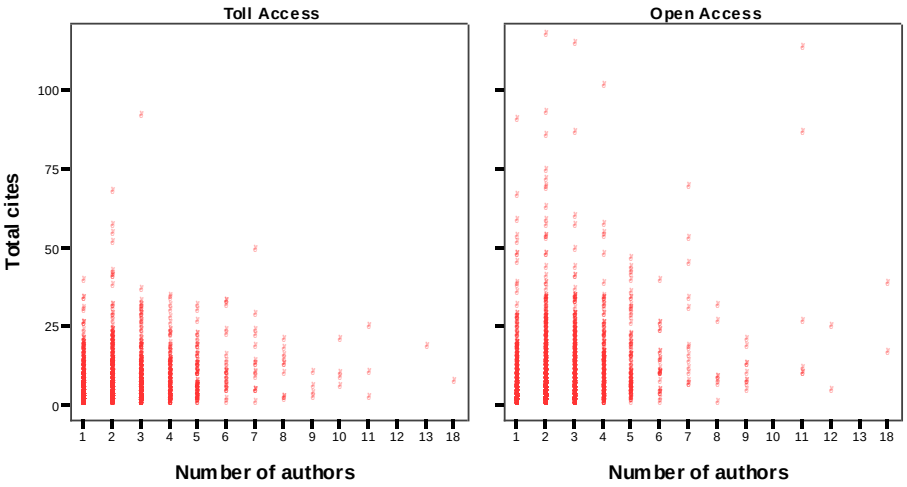


Figure 6.21 OA/TA author citation scatterplots

A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 6.21. For TA articles, this was 0.21 and for OA articles, this was 0.19. Whilst these correlations were significant ($p<0.01$), they indicate a relatively weak correlation. Taking just journal and author self–citations and comparing this total to the level of authorship revealed no substantial differences between the two sets of data; the correlations were 0.309 and 0.288 for OA and TA articles respectively ($p<0.01$). Table 6.7 shows Pearson correlation coefficients when taken at subject level for OA and TA articles against author frequency and citation count.

Table 6.7 Correlation by subject for author/number of citations

	Toll access		Open access	
Subject	Correlation	Significance	Correlation	Significance
Ecology	0.138	0.001	0.157	0.001
Economics	0.022	0.331	-0.021	0.287
Applied maths	0.006	0.451	0.012	0.377
Sociology	0.077	0.010	0.148	0.010

The results from Table 6.7 show a very mixed picture, with OA economics articles, for example, exhibiting a negative correlation, i.e., the greater the number of authors, the fewer the number of citations. Correlations were again weak and in some cases were not statistically significant; overall, the results appear inconclusive.

Correlations between journal impact factor and the number of authors was for all OA articles 0.16 and for all TA articles 0.25. Table 6.8 shows this at subject level. Results were mixed, with negative and positive correlations evident. Apart from ecology, none of the results were statistically significant. Like the results from Table 6.7 the results are also inconclusive, with none of the subjects having a Pearson correlation coefficient above 0.133.

Table 6.8 Correlation by subject for impact factor and numbers of authors

	Toll access		Open access	
Subject	Correlation	Significance	Correlation	Significance
Ecology	-0.112	0.004	-0.133	0.001
Economics	0.013	0.401	-0.041	0.132
Applied maths	-0.008	0.429	0.039	0.154
Sociology	-0.028	0.194	-0.010	0.436

6.7. Search engine success

As a by-product of the data gathering, the success of OAIster, OpenDOAR, *Google* and *Google Scholar* in retrieving OA versions of articles was measured. The search sequence started with OAIster, and proceeded through OpenDOAR, and if necessary *Google Scholar*, and finally *Google*. OAIster and OpenDOAR were always searched; if these two failed to produce a hit, then *Google Scholar* was searched, and finally, if necessary, *Google* was used. Given the search protocol adopted, the results for *Google* and *Google Scholar* cannot be said to reflect the absolute potential of either of them. However, taken together, they jointly found 85.89% of the articles. Figure 6.22 shows the relative success of *Google* and *Google Scholar* compared to OAIster or OpenDOAR. Apart from the combined OAIster and OpenDOAR bar, each of the bars in the chart represents the exclusive hits for that particular search tool. For the combined OAIster and OpenDOAR entry, this is where both of them located the same OA articles.

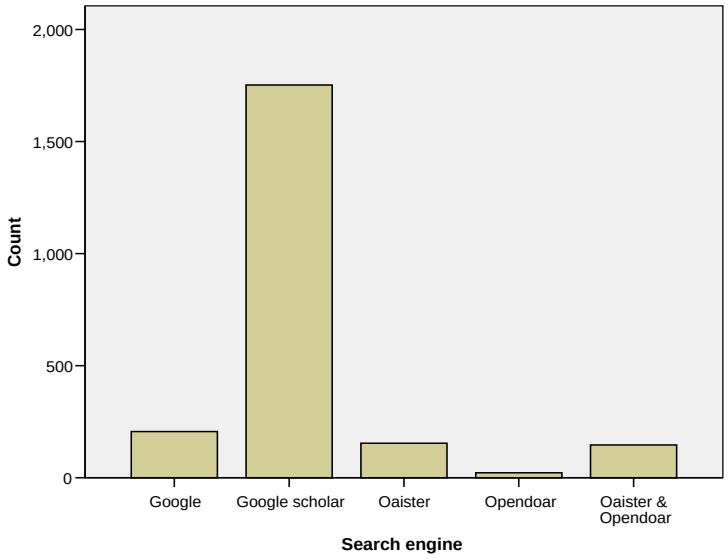


Figure 6.22 Search tool success rate

The percentage of records found for each search tool was; *Google* 9.0%, *Google Scholar* 76.9%, OAIster 6.8%, OpenDOAR 0.9% and where OAIster and OpenDOAR retrieved the same article, their combined score was 6.4%. Table 6.9 gives a more detailed breakdown of search tool hits, by subject.

Table 6.9 Break down of OA hits by subject and search tool

			Subject				
			Ecology	Economics	Applied Maths	Sociology	Total
	Google	Count	98	51	40	17	206
		% within Subject	15.9%	6.9%	5.9%	6.9%	9.0%
	Google scholar	Count	497	531	500	224	1752
		% within Subject	80.4%	71.9%	73.7%	91.4%	76.8%
	Oaister	Count	2	130	21	1	154
		% within Subject	.3%	17.6%	3.1%	.4%	6.8%
	Opendoar	Count	1	4	16	1	22
		% within Subject	.2%	.5%	2.4%	.4%	1.0%
	Oaister & Opendoar	Count	20	23	101	2	146
		% within Subject	3.2%	3.1%	14.9%	.8%	6.4%
Total		Count	618	739	678	245	2280
		% within Subject	100.0%	100.0%	100.0%	100.0%	100.0%

Taken at regional level, the hits shown in Figure 6.23 shows the dominance of *Google Scholar* by count in North America.

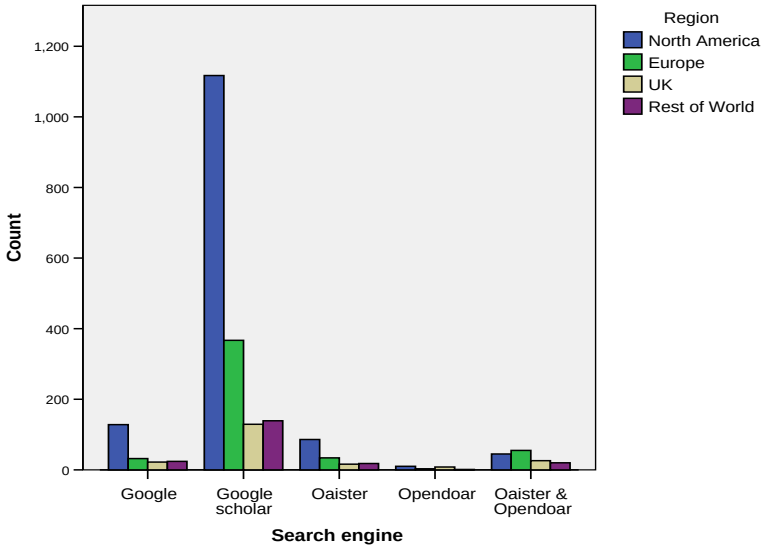


Figure 6.23 OA article hits by region and search tool

Table 6.10 shows in detail the hits by percentage by subject, search engine and territory. Whilst the counts shown in Figure 6.23 show the overall success of *Google Scholar*, Table 6.10 demonstrates that there are regional differences by subject and search engines. In the UK, with the exception of sociology there were a number of hits using OAIster and Open DOAR; similarly in Europe and the Rest of the World, a similar result is evident for economics and applied maths.

Subject				Region: North America	Europe	UK	Rest of World	Total
Ecology		Google	Count	57	15	16	10	98
			% within Region	14.9%	12.4%	24.6%	20.4%	15.9%
		Google scholar	Count	322	100	38	37	497
			% within Region	84.1%	82.6%	58.5%	75.5%	80.4%
		Oaister	Count	2	0	0	0	2
			% within Region	.5%	.0%	.0%	.0%	.3%
		Opendoar	Count	0	0	1	0	1
			% within Region	.0%	.0%	1.5%	.0%	.2%
		Oaister & Opendoar	Count	2	6	10	2	20
			% within Region	.5%	5.0%	15.4%	4.1%	3.2%
	Total		Count	383	121	65	49	618
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%
Economics		Google	Count	38	6	4	3	51
			% within Region	7.3%	6.3%	5.8%	5.5%	6.9%
		Google scholar	Count	390	58	48	35	531
			% within Region	75.0%	61.1%	69.6%	63.6%	71.9%
		Oaister	Count	76	25	14	15	130
			% within Region	14.6%	26.3%	20.3%	27.3%	17.6%
		Opendoar	Count	2	1	1	0	4
			% within Region	.4%	1.1%	1.4%	.0%	.5%
		Oaister & Opendoar	Count	14	5	2	2	23
			% within Region	2.7%	5.3%	2.9%	3.6%	3.1%
	Total		Count	520	95	69	55	739
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%
Applied Maths		Google	Count	19	9	2	10	40
			% within Region	6.5%	3.5%	4.7%	11.4%	5.9%
		Google scholar	Count	233	190	19	58	500
			% within Region	79.5%	74.8%	44.2%	65.9%	73.7%
		Oaister	Count	7	9	2	3	21
			% within Region	2.4%	3.5%	4.7%	3.4%	3.1%
		Opendoar	Count	7	2	6	1	16
			% within Region	2.4%	.8%	14.0%	1.1%	2.4%
		Oaister & Opendoar	Count	27	44	14	16	101
			% within Region	9.2%	17.3%	32.6%	18.2%	14.9%
	Total		Count	293	254	43	88	678
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%
Sociology		Google	Count	14	2	0	1	17
			% within Region	7.4%	9.5%	.0%	10.0%	6.9%
		Google scholar	Count	172	19	24	9	224
			% within Region	90.5%	90.5%	100.0%	90.0%	91.4%
		Oaister	Count	1	0	0	0	1
			% within Region	.5%	.0%	.0%	.0%	.4%
		Opendoar	Count	1	0	0	0	1
			% within Region	.5%	.0%	.0%	.0%	.4%
		Oaister & Opendoar	Count	2	0	0	0	2
			% within Region	1.1%	.0%	.0%	.0%	.8%
	Total		Count	190	21	24	10	245
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%

Table 6.10 Search tool success by subject and region

6.8. Impact factor

Eighty percent of all impact factors for the 64 journals lie in the range 0.9 to 4.3 with a mean of 2.5. There are two extreme outliers, one at 7.2 (applied maths) and the other at 14.9 (ecology); these two account for 91 (1.9%) of all the articles. When examined by subject, there are noticeable differences in the range and concentration of impact factor scores. The boxplot shown in Figure 6.24 allows a comparison of the four subjects.

Asterisks indicate outliers that are more than three box-lengths away from the box, circles show outliers which are more than 1.5 box lengths away from the box.

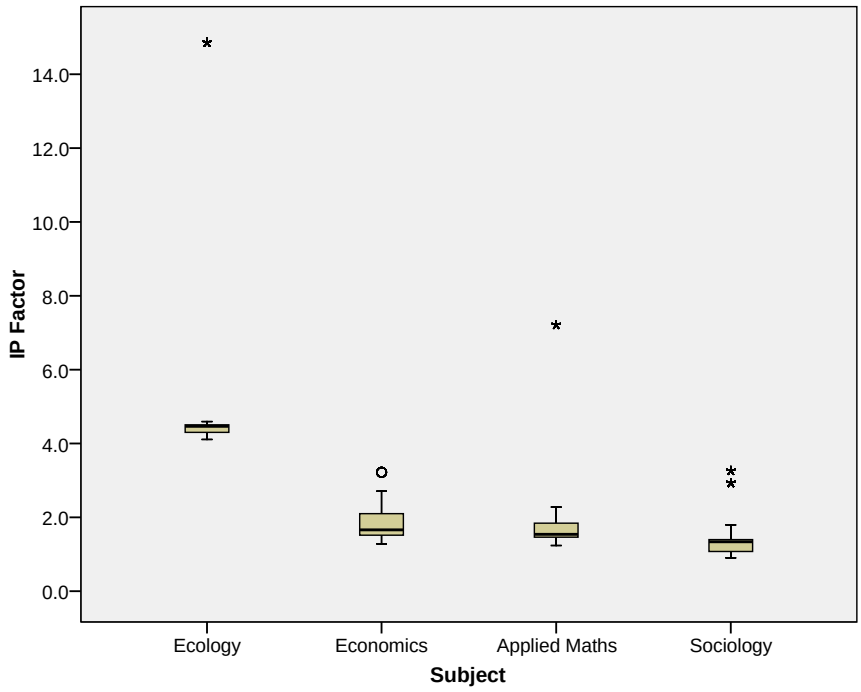


Figure 6.24 Boxplot of impact factor scores by subject.

Three further outliers are identified in Figure 6.24, one for economics and two for sociology. Discounting the outliers, the impact factor ranges for the selected journals are narrow; this is as expected given the selection was purposely taken from high impact journals. Noticeable is the relative position of the subjects, with ecology (a typical science subject) having journals with a higher impact factor than sociology. Outlier details are given in the Table 6.11.

Table 6.11. Impact factor outlier details

Journal	Subject	Impact Factor	Number of articles
American Journal of Sociology	Sociology	2.933	58
American Sociology Review	Sociology	3.262	76
Journal of Economic Geography	Economics	3.222	18
Siam Review	Applied maths	7.273	20
Trends in Ecology & Evolution	Ecology	14.864	71

The boxplot shown in Figure 6.25 shows that overall, the range of journal impact factors is higher for OA articles than TA articles, suggesting that OA articles are more likely to be found in higher impact factor journals. A Chi-squared test on the association of impact factor and OA status of an article ($\chi^2(62) = 855.549$, $p < .001$) confirms a significant association at the higher levels of impact factor; the higher the impact factor of a journal, the more likely that some of its articles will be OA.

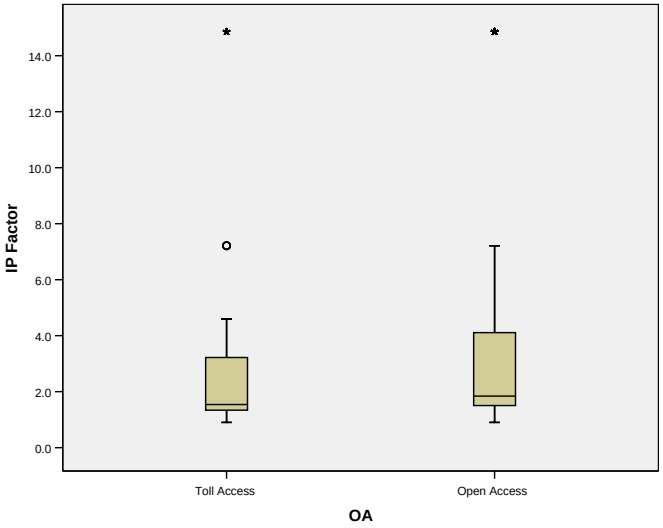


Figure 6.25 Impact factor by OA/TA status

The boxplot at Figure 6.26 breaks down the relationship shown in Figure 6.25 into subjects. It becomes apparent that the result in Figure 6.25 is influenced by the results for economics and sociology, which have a greater number of its OA articles in higher impact journals. For the other two subjects, this is only marginally the case. It is noticeable however that there only seven journals in the ecology sample within a relatively tight range of impact factors with the exception of an outlier at 14.9, see Appendix B and Table 6.11 for details.

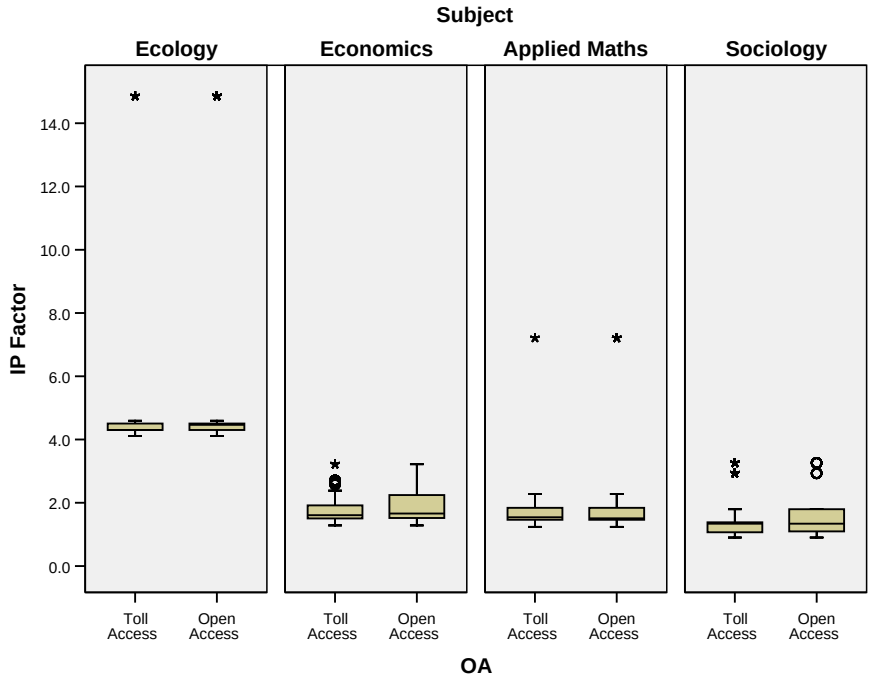


Figure 6.26 Impact factor by OA/TA status and subject

A Chi-squared test of association by impact factor and OA status of an article for each of the four subjects was carried out. For ecology ($\chi^2(6) = 42.093$, $p < .001$), economics ($\chi^2(21) = 197.768$, $p < .001$), applied maths ($\chi^2(15) = 34.3$, $p < .003$) and for sociology ($\chi^2(18) = 111.921$, $p < .001$). The results confirm a modest, but significant association at the higher levels of impact factor; the higher the impact factor of a journal, the more likely that any of its articles will be OA. Given the results from the boxplot at Figure 6.26, the association is only very weak for ecology and applied maths.

6.9. Within journal comparisons

Figure 6.27 to Figure 6.30 illustrate the split between OA and TA articles within the journals for each subject.

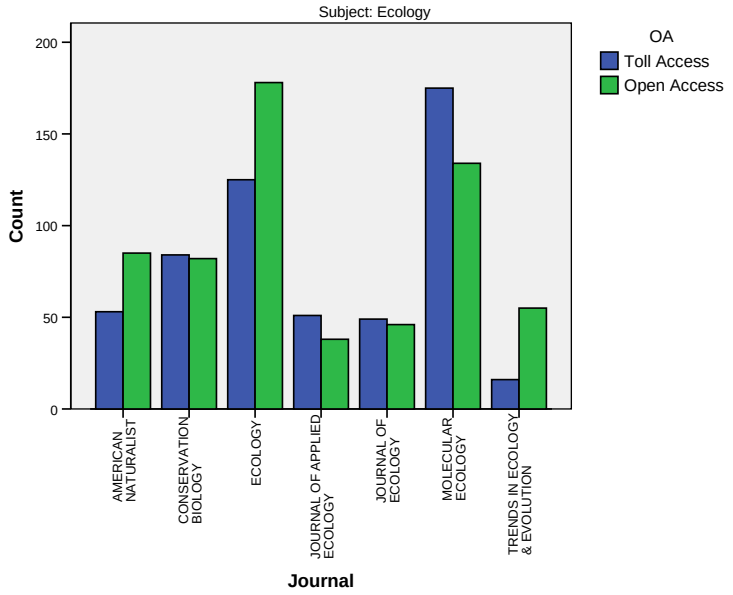


Figure 6.27 OA/TA article split for ecology

The split for ecology is roughly even between OA and TA articles. A noticeable exception is the high impact journal *Trends in Ecology and Evolution*; although this has the lowest number of articles, 77.5% are OA. Overall 52.8% of ecology articles are OA.

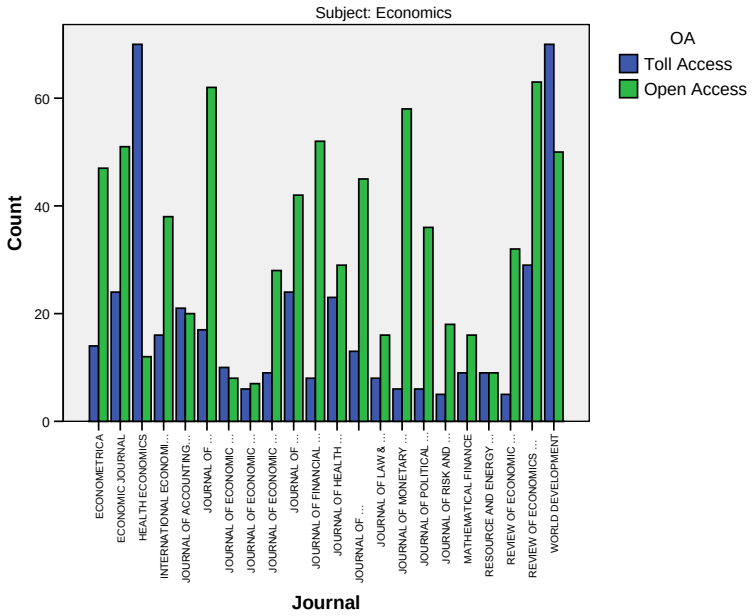


Figure 6.28 OA/TA article split for economics

The majority of articles in the selected economics journals are OA, with only four journals having more of its articles TA. The *Journal Health Economics* has the fewest with only 12 out of its 82 articles being OA. *The Journal of Monetary Economics* had

58 of its articles OA out of its total of 64. Economics has at, 64.8% the highest rate of OA amongst the four subjects.

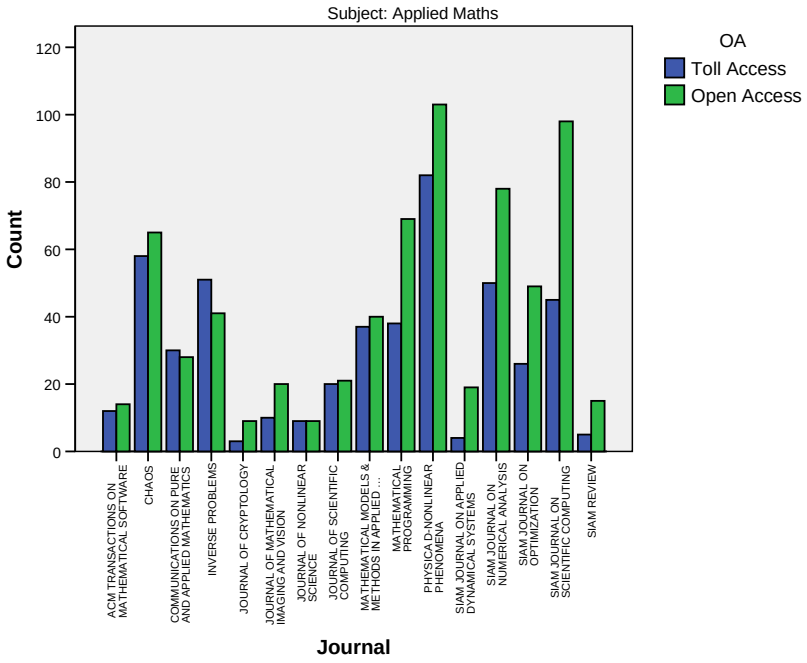


Figure 6.29 OA/TA article split for applied maths

For applied maths, the split between OA and TA article counts is much more closely matched with three journals having more TA than OA articles. Overall, 58.5% of articles were OA.

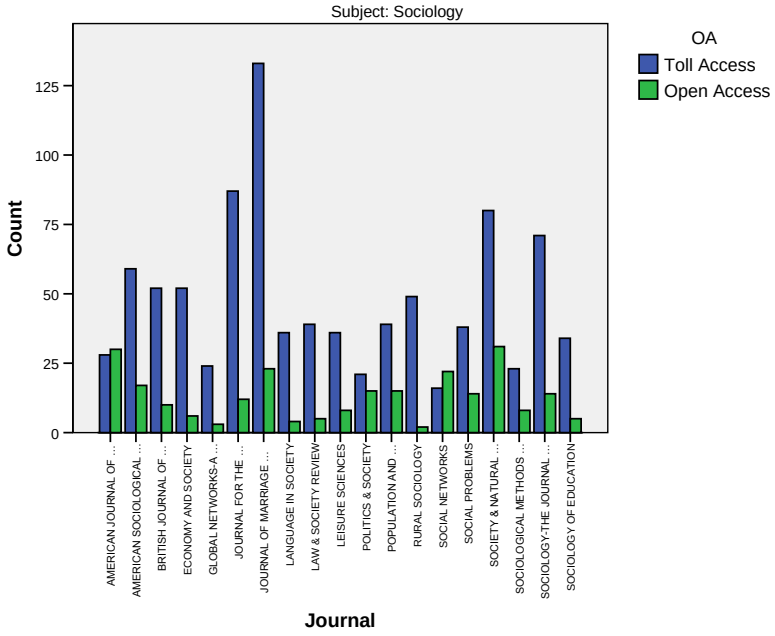


Figure 6.30 OA/TA article split for sociology

Sociology has far fewer OA articles with only two journals having more OA than TA articles; these are the *American Journal of Sociology* and *Social Networks*, both of which have American publishers. Overall, sociology has the lowest rate of OA at 21.1%.

6.10. Distribution of citations within subjects

The distributions of citations within each subject and by OA/TA status is skewed in favour of a relatively small number of articles, which receive the majority of citations. Table 6.12 gives the overall distribution of citations by article count, by subject and OA/TA status.

Table 6.12 Distribution of all citations by percentage article count

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All articles	4633	6.7	15.6	29.4	54.6
Toll Access	2353	7.3	16.3	29.9	54.3
Ecology	553	10.5	22.9	40.8	67.2
Economics	402	8.7	18.4	32.8	55.9
Applied Math	480	7.5	16.0	30.0	53.5
Sociology	918	7.2	15.8	29.4	53.3
Open Access	2280	6.9	16.1	30.2	55.8
Ecology	618	9.2	21.8	40.0	67.6
Economics	739	6.4	15.3	29.4	55.1
Applied Math	678	7.4	16.8	31.3	55.9
Sociology	245	8.2	16.7	29.8	55.9

The same article records are shown below in Table 6.13, but without any self-citations, that is it records citations from other authors only.

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All articles	4633	5.5	13.3	25.8	49.4
Toll Access	2353	6.3	14.3	26.7	49.4
Ecology	553	9.2	20.8	37.9	65.0
Economics	402	7.2	15.9	29.3	51.7
Applied Math	480	5.6	12.5	23.9	44.4
Sociology	918	6.3	14.0	26.8	49.8
Open Access	2280	5.5	13.6	26.5	50.4
Ecology	618	7.4	18.4	35.6	64.0
Economics	739	5.5	13.5	26.7	51.8
Applied Math	678	5.7	12.5	24.9	46.7
Sociology	245	7.3	15.5	27.7	53.5

Table 6.13 Distribution of other author only citations by percentage article count

Taking a sample journal from each subject and splitting these by their OA/TA status shows that the characteristics exhibited in Table 6.12 are approximately apparent at single journal level as well. The four journals selected were:

Subject	Title
Ecology	*American Naturalist*
Economics	*World Development*
Applied Maths	*Chaos*
Sociology	*Journal of Marriage and the Family*

Table 6.14 Journal titles selected for citation distribution

Table 6.15 shows the distribution of all citations for each of the selected journals by their subject. The OA sociology sample is based on a relatively small number of articles.

Table 6.15 Distribution of all citations by percentage article count at journal level

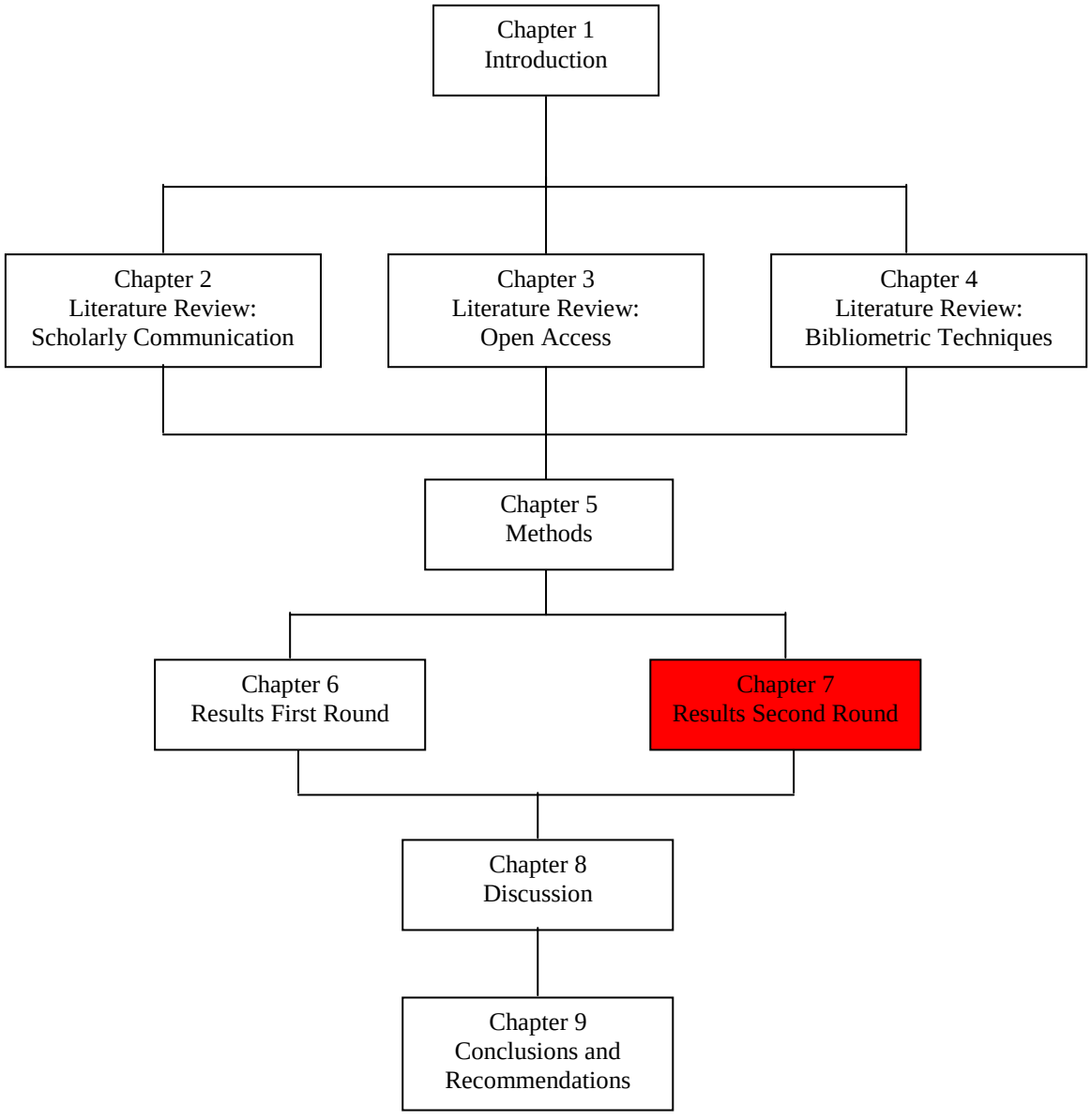
Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
Toll Access					
Ecology	53	9.6	21.2	40.4	69.2
Economics	70	11.4	24.3	41.4	67.1
Applied Math	58	8.6	17.2	32.8	60.3
Sociology	87	8.0	16.1	29.9	52.9
Open Access					
Ecology	85	11.8	24.7	41.2	68.2
Economics	50	5.7	12.9	24.3	42.9
Applied Math	65	12.3	24.6	41.5	66.2
Sociology	23	8.7	21.7	34.8	69.6

The same article records are shown below in Table 6.16 but without any self-citations, that is it counts citations from other authors only.

Table 6.16 Distribution of author only citations by percentage article count

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
Toll Access					
Ecology	53	7.5	16.9	33.9	66.0
Economics	70	10.0	22.8	38.5	64.2
Applied Math	58	5.2	12.9	27.6	53.4
Sociology	133	7.5	16.5	29.3	54.1
Open Access					
Ecology	85	10.0	22.3	38.8	67.0
Economics	50	8.0	18.0	34.0	60.0
Applied Math	65	7.7	18.4	33.9	55.4
Sociology	23	8.7	19.4	34.7	60.9

Chapter 7 Results: Second Round Data Collection



7.1. Introduction

A second round of data collection was undertaken in line with Objectives 2-5 as given in Chapter 1. The results are reported in this chapter.

7.2. Objective 2

Objective 2 takes the subject examined in Chapter 6 with fewest OA articles (sociology), and takes a second sample of articles from high impact journals in order to establish whether the OA advantage identified for the subject can be replicated in articles published approximately one year later in 2004. Appendix D gives details of the journal titles selected, their impact factor and the number of articles collected from each. The total number of article records collected was 931. Of these, 227 (24.4%) were OA, leaving the remaining 704 (75.6%) as TA. In total, the articles accrued 2450 citations between them; 257 articles did not receive any citations at all.

7.3. Distribution of citation counts

Figure 7.1 illustrates the positive skew of citations when all of the 931 article citation records for both OA and TA records are counted and plotted. The data has a mean of 2.6, a standard deviation of 3.5, and a mode of 0 and a median of 2. The citations counts are distributed such that 82.5% of all citations fall between 0 and 4. The overall range for citation counts is 0-35.

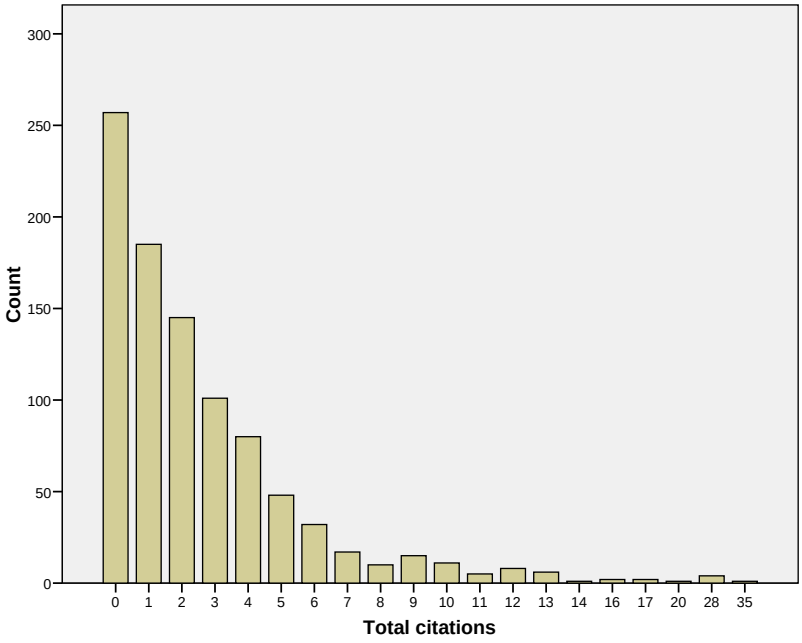


Figure 7.1 Distribution of all citations

The line graph at Figure 7.2 compares the citation counts for both the OA and TA articles. Differences are noticeably apparent at the zero and five-citation counts. For TA articles, 67.6% of citations fall between 0-2, whereas for OA articles, 48.9% of citations fall in this range. Of the 257 articles that did not attract any citations, the majority (213) were TA (82.9%).

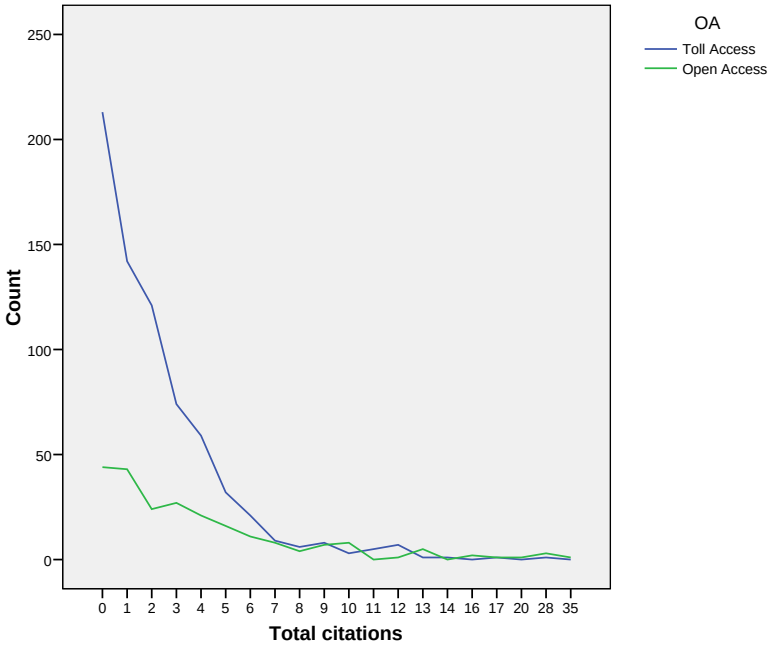


Figure 7.2 OA and TA citation distribution article

Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 3.96 compared to 2.2 for the TA articles. This gives a citation advantage in favour of OA articles of 80% (OA-TA/TA citation counts *100). The OA advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals. When these citations were excluded, the mean citation counts for the two article sets were OA 2.79 and TA 1.44. This extends the OA advantage, which becomes 94%.

A two sample independent *t*-test and a two sample Mann-Whitney test was carried out to test whether there was a statistically significant difference between the mean citation counts of OA and TA articles. In all cases, there was a significant difference between the means for both tests, indicating that the two populations of citation counts are drawn from two different populations ($p<0.001$).

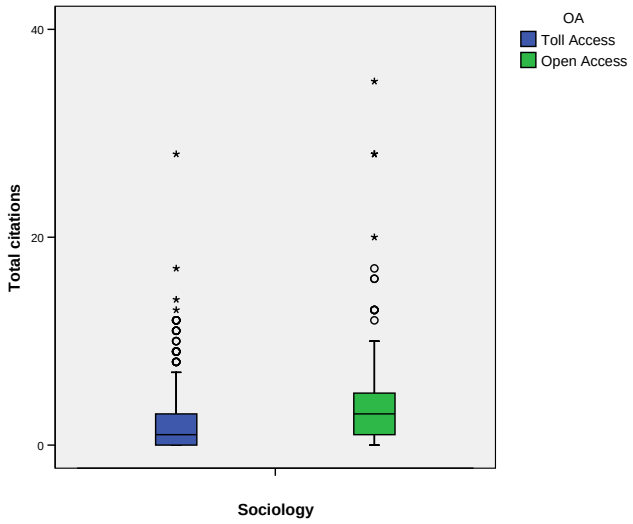


Figure 7.3 Boxplot of the distribution of citations

The boxplot at Figure 7.3 helps illustrate the results from the *t*-test and the Mann-Whitney, showing that OA articles consistently show a greater median citation value than TA articles. Asterisks indicate outliers that are more than three box-lengths away from the box, and circles show outliers which are more than 1.5 box lengths away from the box. The boxplot in Figure 7.4 shows the distribution of other author citations by OA status. The median value for OA citations is greater as is the range.

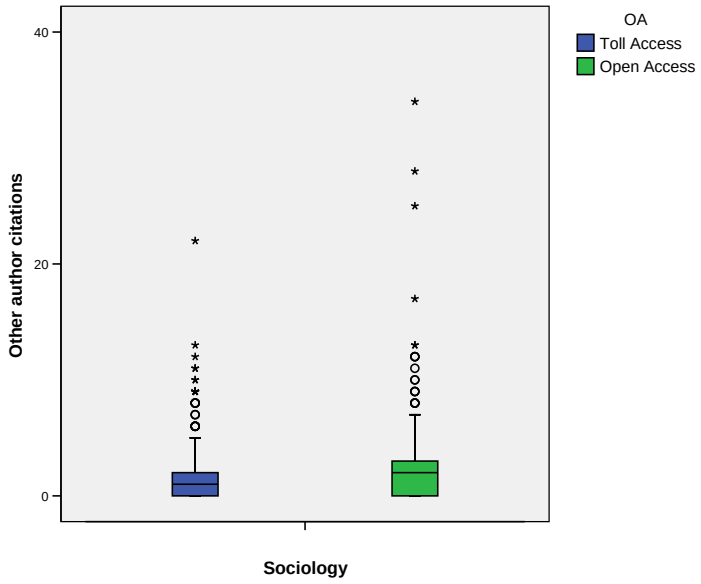


Figure 7.4 Boxplot for other author citations

7.4. Self-citation counts

The rate of self-citation varies between OA and TA articles. Figure 7.5 shows by percentage article count the distribution of all categories of self-citation for both OA and TA articles. The data for OA articles has a mean of 1.16 and a standard deviation of 1.847 with a modal value of 0 and a median of 1. For the TA data, the articles have a mean of 0.77 and a standard deviation of 1.26 with a modal value of 0 and a median of 0.

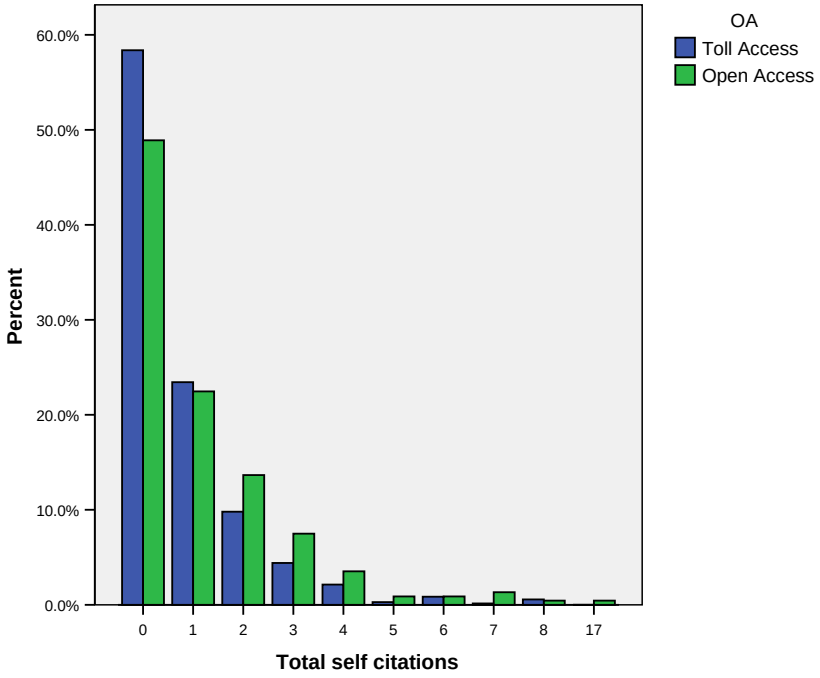


Figure 7.5 Distribution of self-citations by OA/TA status

Figure 7.6 shows the relative closeness of the self-citation counts when they are taken together. However, the most noticeable difference is apparent at the zero count where 58.4% of TA articles have no self-citations, whereas for OA articles this is 48.9%.

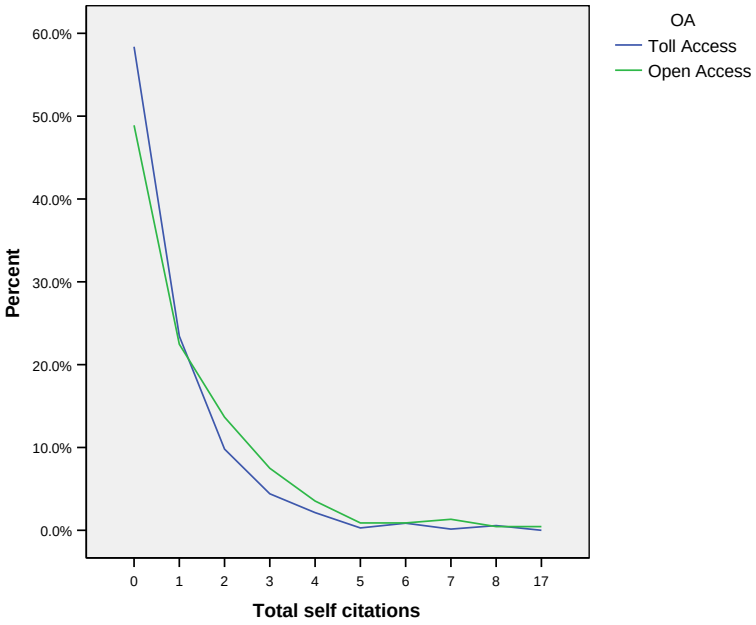


Figure 7.6 All OA and TA self citations

Figure 7.7 show a breakdown of the gross citation count by the four types identified, three of which are related to author or journal self-citation.

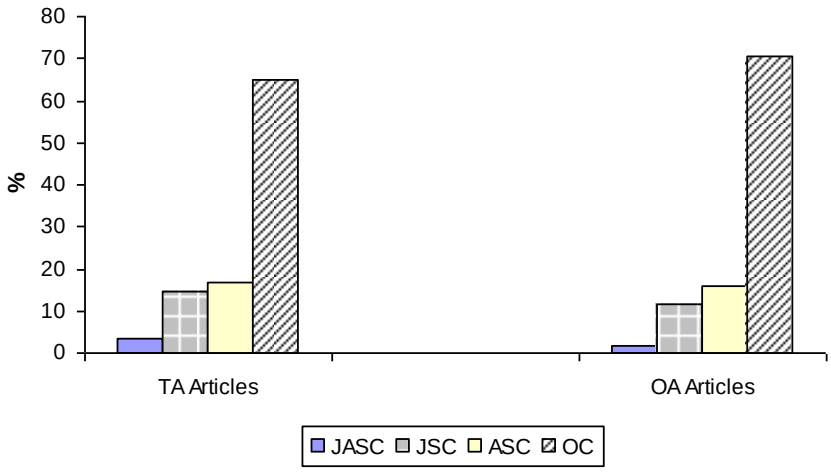


Figure 7.7 Breakdown of TA/OA citations

The combined self-citation rate for TA articles is 35% and for OA articles, this is 29%. Figure 7.8 shows a comparison between each citation category and their OA/TA status and

illustrates that OA articles have a greater percentage of their citations in the 'other citation' category than TA articles. Conversely, TA articles have consistently more citations in the self-citation categories.

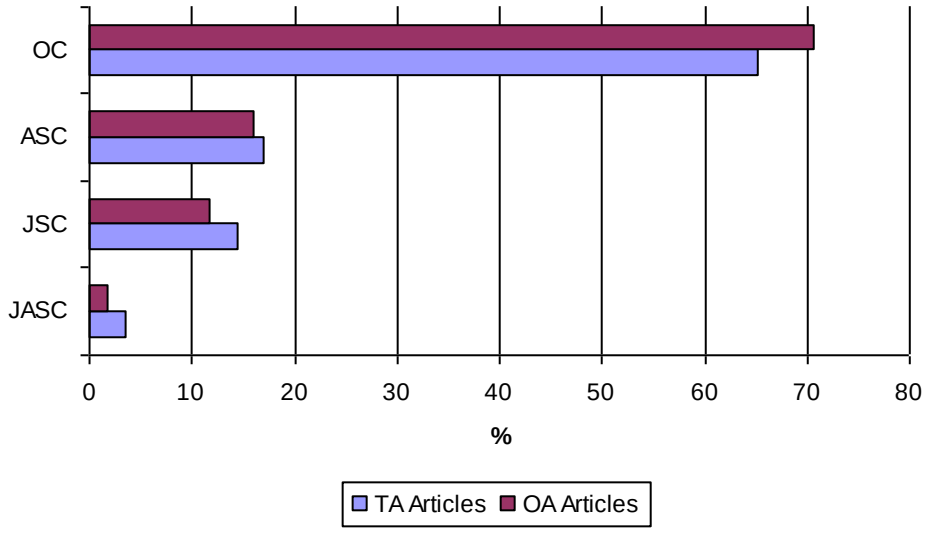


Figure 7.8 Articles by their citation category

However, OA articles consistently have a higher individual citation count for the articles within the self-citation categories outlined above, despite there being fewer OA articles than TA articles in these self-citation categories. The mean number of journal and author self-citations for OA articles was 1.16, and for TA articles, this was 0.77. These means were compared using the independent 2 sample *t*-test; the result showed them to be from populations with different means ($p<0.001$).

The scatter plot shown in Figure 7.9 shows the relationship between other author citations (OC) and all types of self-citation. A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.9. The result was significant ($p<0.01$). For OA articles, this was 0.243 and for TA articles it was 0.294.

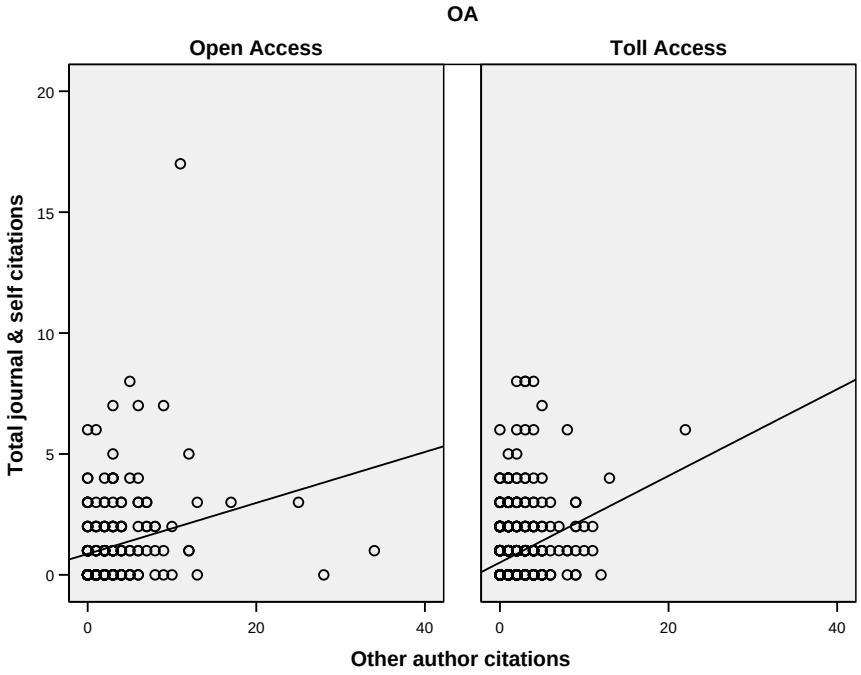


Figure 7.9 OA/TA scatterplot of self-citations to other author citations

7.5. Author frequency and OA/TA status

The mean number of authors for OA articles was 1.93 and for TA articles, this was 1.81. Figure 7.10 illustrates the distribution of author counts by the OA/TA status of their articles. The distribution is heavily skewed by the predominance of TA articles and single authorship.

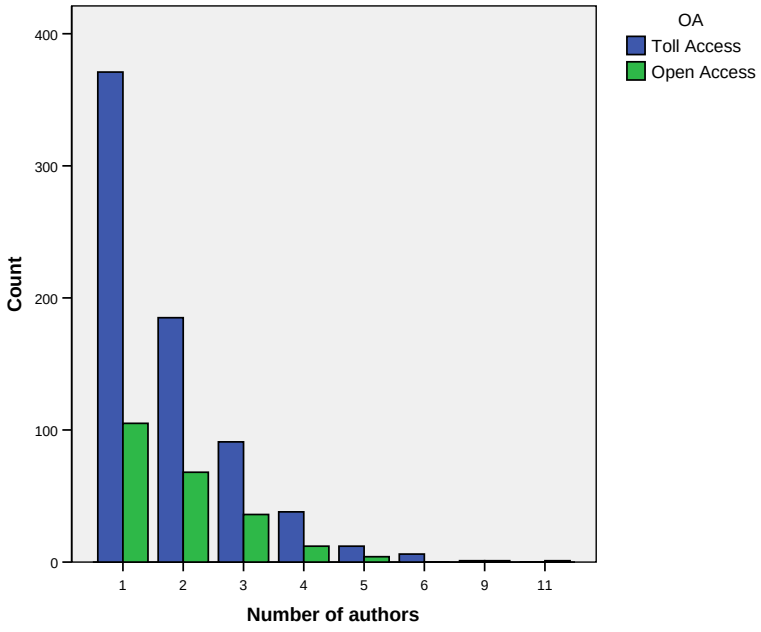


Figure 7.10 Articles by author count and OA/TA status.

There are noticeably more single authored TA (371) articles than there are OA articles (105). In every case, TA articles outnumber OA articles by author count. Given the low rate of OA in sociology this is an unsurprising result. The results from a Chi-square test showed there was no association overall between the number of authors and the OA/TA status of an article. The 931 articles sampled yielded by first author affiliation 43 countries of origin (six articles had no first author affiliation given). For analysis purposes, these were grouped into four regions: North America, continental Europe, UK and the Rest of the World.

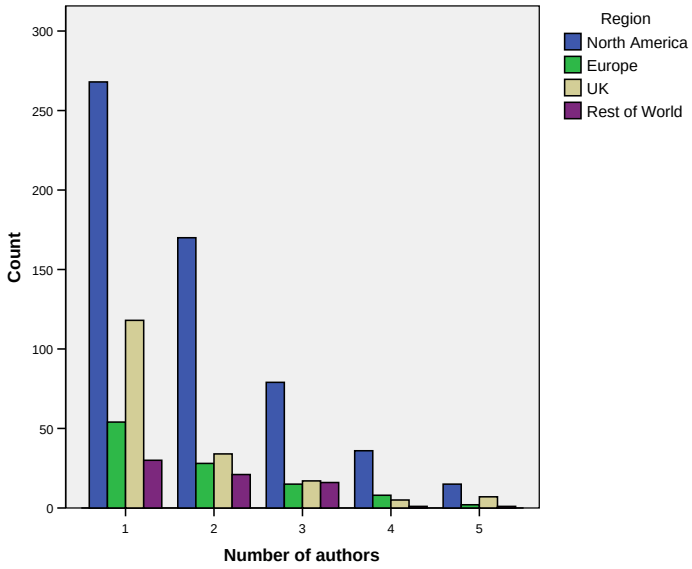


Figure 7.11 Articles by author count and region.

Figure 7.11 shows the split when author counts are limited to no more than five, six and greater account for 1.0% of articles across all four regions. North America predominates with 568 articles followed by the UK with 181 articles, Europe with 107 articles and the Rest of the World with 69.

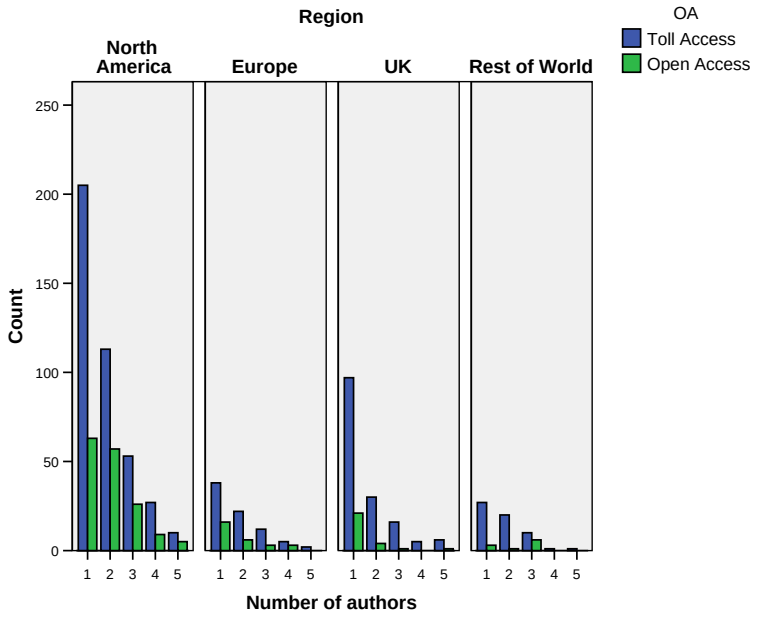


Figure 7.12 Author count by region and OA status

Figure 7.12 shows the frequency of TA articles in every region irrespective of author count and the dominance of the USA in the number of articles published; it is a similar result to that found in the first round of data collection for this subject.

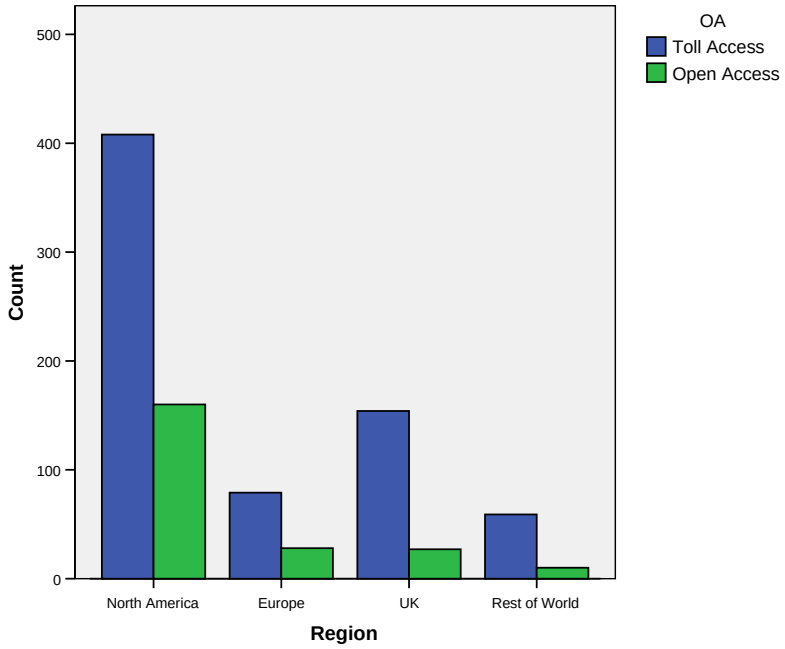


Figure 7.13 Number of OA/TA articles by region

Figure 7.12 illustrates the dominance of TA articles, with the Rest of the World and the UK having the highest percentage of TA articles at 85% and 86% respectively. North America has 70% of its articles TA and Europe has 74%.

7.6. Correlations

The pair of scatterplots in Figure 7.14 shows the distribution of citations against the number of authors for TA and OA articles.

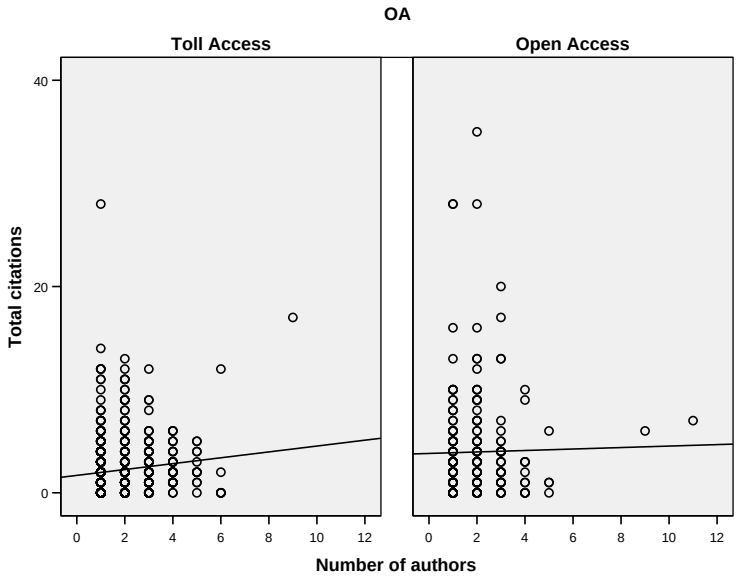


Figure 7.14 OA/TA author citation scatterplots

A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.14. For TA articles, this was 0.114 and for OA articles, this was 0.018; however, the OA result was not significant. Taking just journal and author self–citations and comparing this total to the level of authorship revealed no substantial differences between the two sets of data; the correlations were 0.101 and 0.115 for OA and TA articles respectively. Again, however, the OA result was not significant. Correlations between journal impact factor and the number of authors were also not significant.

7.7. Search engine success

Details of the how the search tools were used is given in section 6.7. Figure 7.15 shows the relative success of *Google* and *Google Scholar* as compared to OAIster or OpenDOAR. Apart from the combined OAIster and OpenDOAR bar, each of the bars in the chart

represents the exclusive hits for that particular search tool. For the combined OAIster and OpenDOAR entry this is where both of them located the same OA articles.

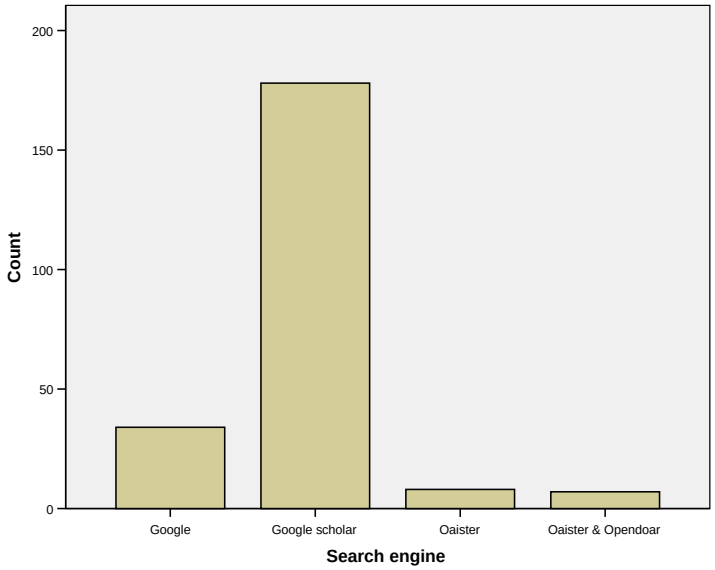


Figure 7.15 Search tool success rate

The percentage of records found for each search tool was; *Google* 14.9%, *Google Scholar* 78.4% (combined score 93.4%), OAIster 3.5%, OpenDOAR found no records exclusively and where OAIster and OpenDOAR retrieved the same article their combined score was 3.1%. If compared to the first round results, although not an absolute like for like match, there is a move away from the dominance of *Google* and *Google Scholar*. Table 6.9 shows the first round results by discovery tool and there is a marginal difference in the score for sociology, which was 91.4% for *Google Scholar* and 6.9% for *Google* with a combined score of 98.3.

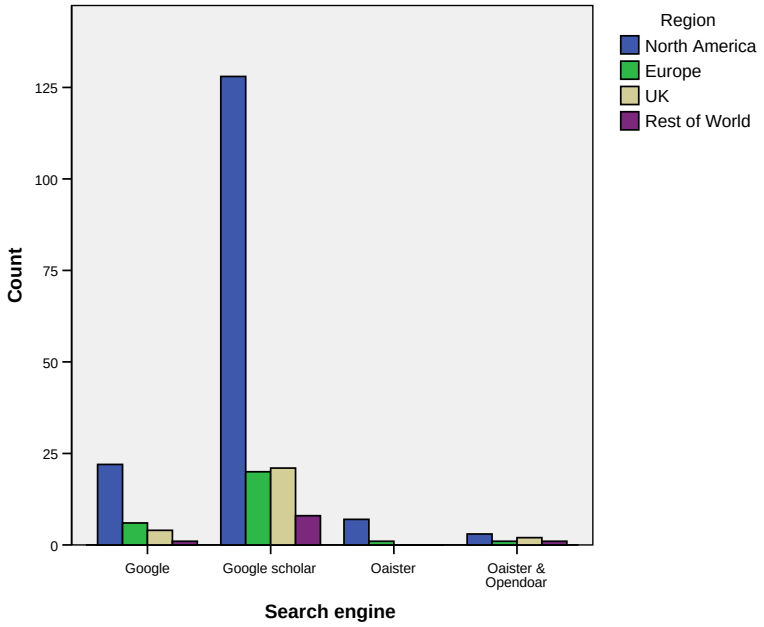


Figure 7.16 OA article hits by region and search tool

Taken at regional level, as in the first round data, the hits in Figure 7.16 show the predominance of *Google* and *Google Scholar* in North America.

		Region				
		North America	Europe	UK	Rest of World	Total
Google	Count	22	6	4	1	33
	% within Region	13.8%	21.4%	14.8%	10.0%	14.7%
Google scholar	Count	128	20	21	8	177
	% within Region	80.0%	71.4%	77.8%	80.0%	78.7%
Oaister	Count	7	1	0	0	8
	% within Region	4.4%	3.6%	.0%	.0%	3.6%
Oaister & Opendoar	Count	3	1	2	1	7
	% within Region	1.9%	3.6%	7.4%	10.0%	3.1%
Total	Count	160	28	27	10	225
	% within Region	100.0%	100.0%	100%	100.0%	100%

Table 7.1 Search tool success by subject and region

Table 7.1 shows in detail the hits by percentage by search engine and territory. Whilst the counts shown in Figure 7.16 show the overall success of *Google Scholar*, Table 7.1 demonstrates that there are some very small regional differences albeit there is a small move away from *Google* and *Google Scholar* when compared to the first round data.

7.8. Impact factor

The boxplot shown in Figure 7.17 indicates that the range of journal impact factors is higher for OA articles than TA articles, suggesting that OA articles are more likely to be found in higher impact factor journals. Asterisks indicate outliers that are more than three box-lengths away from the box, circles show outliers which are more than 1.5 box lengths away from the box.

A Chi-squared test on the association of impact factor and OA status of an article ($\chi^2(26) = 90.486$, $p < .001$) confirms significant association at the higher levels of impact factor; the higher the impact factor of a journal, the more likely that any of its articles will be OA.

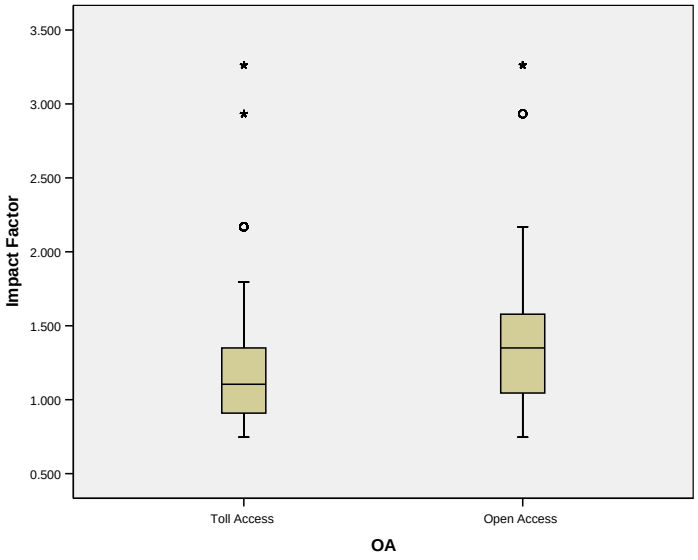


Figure 7.17 Impact factor by OA/TA status

7.9. Within journal comparisons

Figure 7.18 illustrates the split between OA and TA articles within the journals for this sample of sociology.

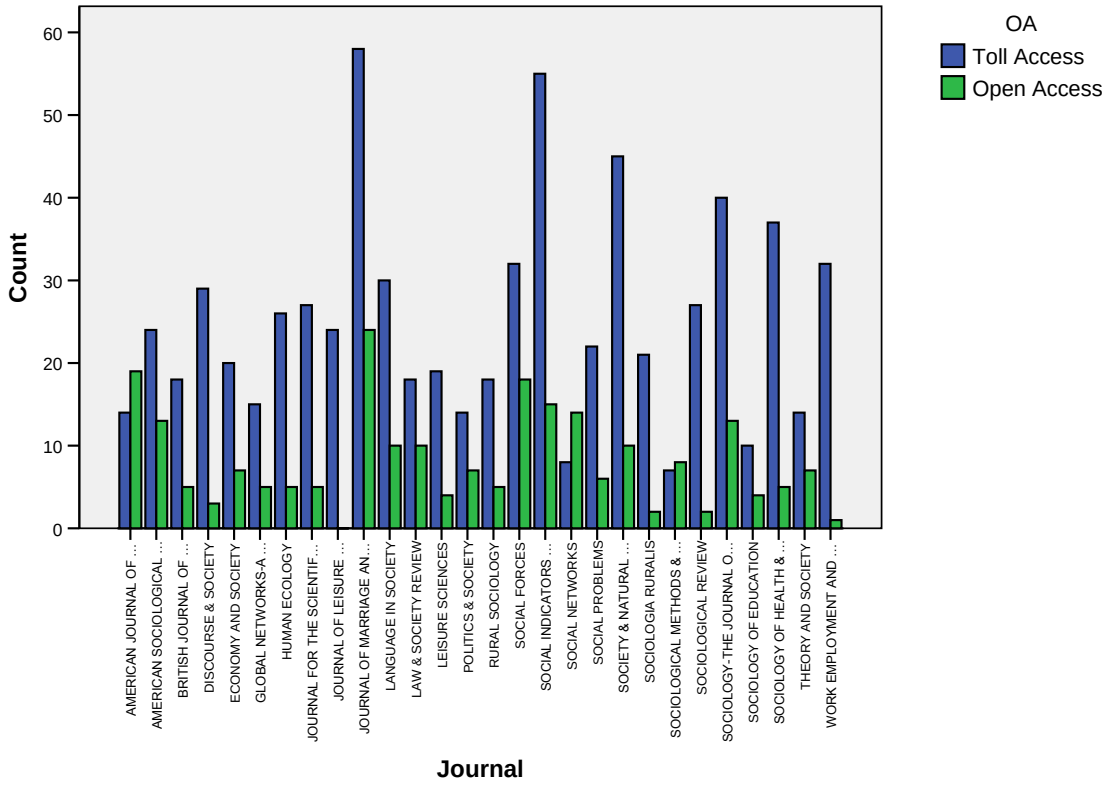


Figure 7.18 OA/TA article spilt by journal title

Sociology has far fewer OA articles with only three journals having more OA than TA articles; these are the *American Journal of Sociology, Social Networks* and *Sociological Methods and Research,* all three of which have American publishers.

7.10. Distribution of citations within subjects

The distributions of citations within this subject and by OA/TA status is skewed in favour of a relatively small number of articles, which receive the majority of citations. Table 7.2 gives the overall distribution of citations by article count, OA/TA status, and also with and without self-citations.

Table 7.2 Distribution of citations by percentage article count

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All citations					
All articles	931	6.2	14.8	27.7	49.3
Toll Access	704	6.7	15.9	28.6	48.7
Open Access	227	6.6	14.9	29.0	51.5
Without self-citations					
All articles	931	4.8	12.2	23.4	43.1
Toll Access	704	5.7	13.5	24.8	43.7
Open Access	227	4.9	13.0	24.7	46.2

Taking the article records from just one journal the *Journal of Marriage and the Family*, and splitting these by their OA/TA status shows that the characteristics exhibited in Table 7.2 are apparent at single journal level as well. Table 7.3 shows the distribution of citations for the *Journal of Marriage and the Family* both with, and without self-citations. At 82 articles, this is a relatively small sample.

Table 7.3 Distribution of citations by percentage article count at journal level

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All citations					
All articles	82	8.5	18.2	31.7	56.1
Toll Access	58	9.5	18.9	32.7	58.6
Open Access	24	6.25	16.6	33.3	54.0
Without self-citations					
All articles	82	8.5	18.3	32.9	52.4
Toll Access	58	8.6	18.9	32.7	53.4
Open Access	24	8.3	16.6	29.2	50.0

7.11. Objective 3

Objective 3 was concerned with ascertaining whether the OA/TA citation advantage is randomly evident in a population of journal articles and whether there is an early access advantage evident from OA articles in terms of patterns of earlier citations.

7.12. Data overview

For this objective, a random sample of articles was taken from the 10,119 articles that appeared in the 112 ecology journals listed in the *Journal Citation Report* for 2005. The sample comprised of 630 article records taken from 82 of the journals, with a sample range of articles of 1-36 with a rounded mean of 8 articles from each journal. Of the 630 articles, 216 (34%) were OA and the remaining 414 (66%) were TA. In total, the articles accrued 4785 citations; 59 articles did not receive any citations at all.

7.13. Distribution of citation counts

Figure 7.19 illustrates the positive skew of citations when all 630 article citation records for both OA and TA records are counted and plotted. The data has a mean of 7.6, a standard deviation of 8.96 with a mode of 1 and a median of 5. The citations counts are distributed such that 78.7% of all citations fall between 0 and 11. The overall range for citation counts was 0-78.

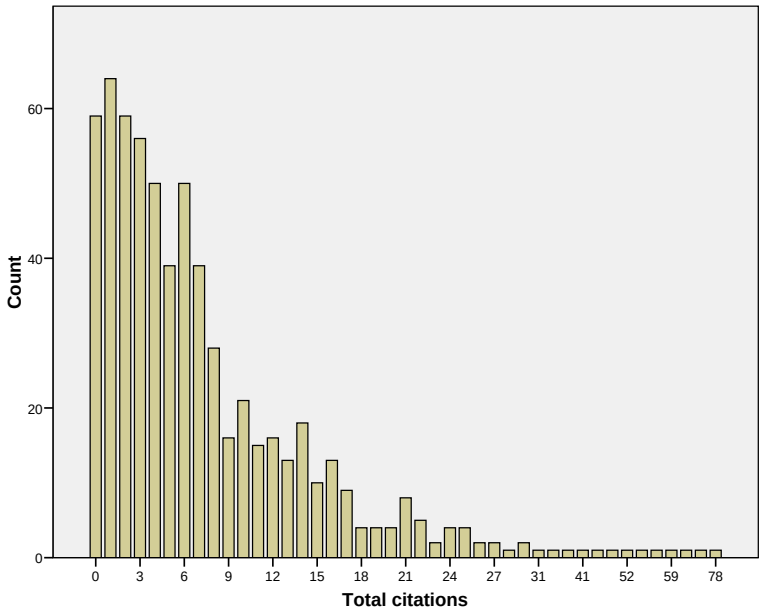


Figure 7.19 Distribution of all citations

The line graph at Figure 7.20 compares the citation counts for all OA and TA articles. Differences are apparent between the zero and six-citation counts, for TA articles 70.5% of citations fall between 0-6, whereas for OA articles 39.4% of citations fall in this range. It is noticeable that 51 of the 58 zero citations counts are TA.

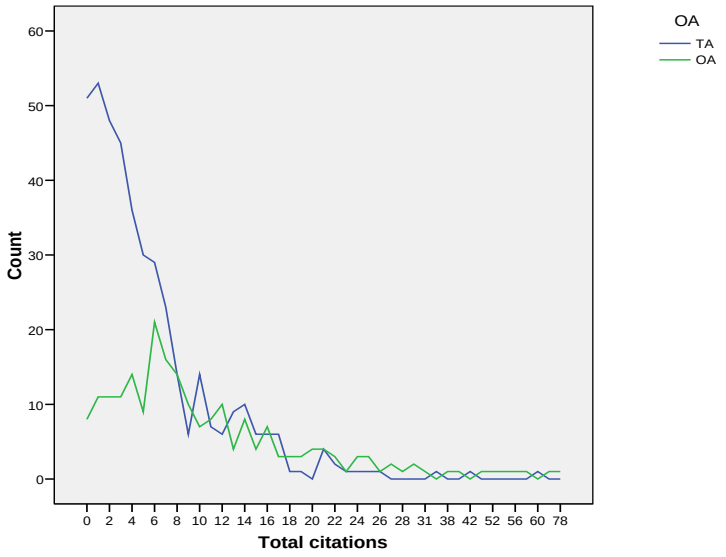


Figure 7.20 OA and TA citation distribution

Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 11.46 compared to 5.58 for the TA articles. This gives a citation advantage in favour of OA articles of 105% (OA-TA/TA citation counts *100). The OA

advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals. When these citations were excluded, the mean citation counts for the two article sets were OA 8.03 and TA 3.41. This extends the OA advantage, which becomes 135%.

A two sample independent *t*-test and a Mann-Whitney test were carried out to test whether there was a statistically significant difference between the mean citation counts of OA and TA articles. In all cases, there was a significant difference between the means for both tests indicating that the two populations of citation counts are drawn from two different populations ($p<0.001$).

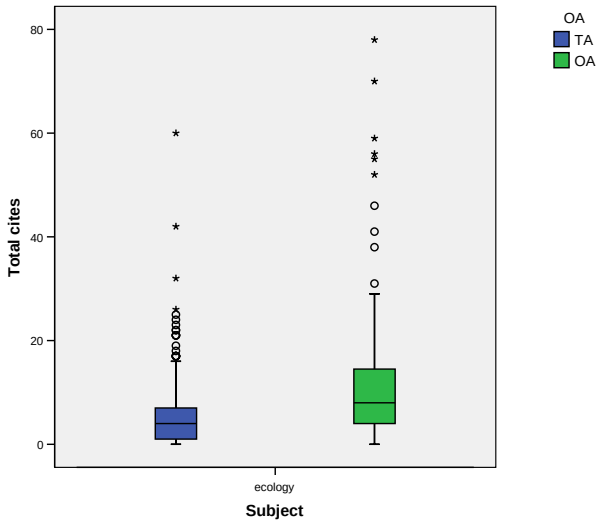


Figure 7.21 Boxplot of the distribution of all citations

The boxplot at Figure 7.21 illustrates the results from the *t*-test and the Mann-Whitney showing that of OA articles consistently show having a greater median citation value than TA articles. Asterisks indicate outliers that are more than three box-lengths away from the box, circles show outliers which are more than 1.5 box lengths away from the box. The boxplot in Figure 7.22 shows the distribution of other author citations by OA status. The range and spread of OA citations is noticeably greater than the TA citations.

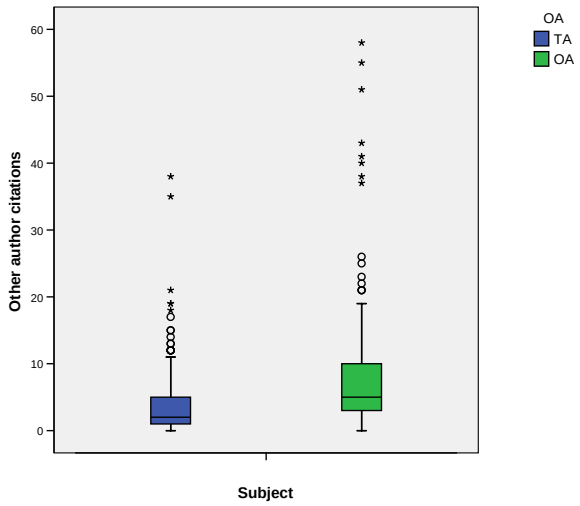


Figure 7.22 Boxplot of other author citations

Figure 7.23 below shows the incidence of OA articles across the range of impact factors by their mean citation advantage although given the sample is random not every impact factor has both a OA and TA article presence although most do.

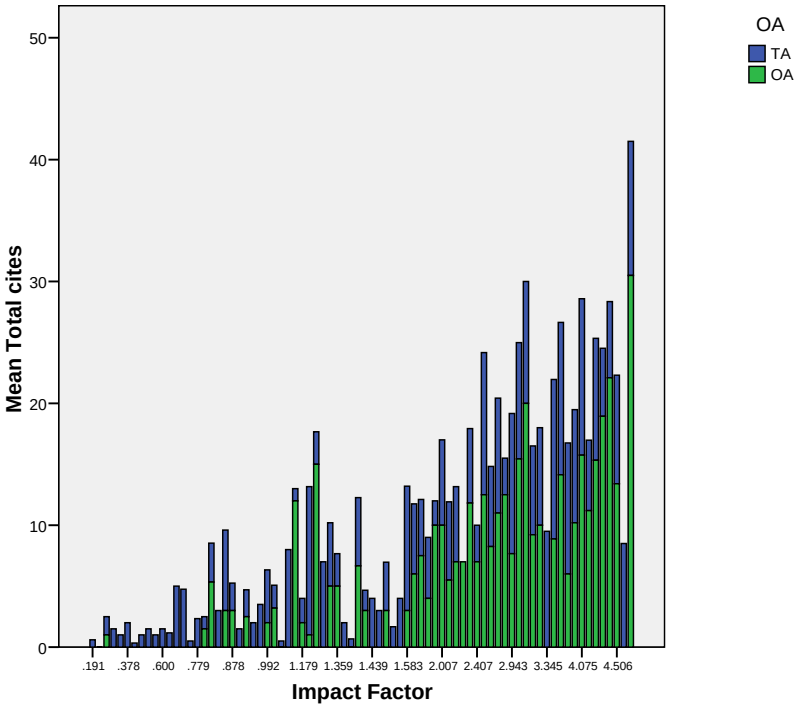


Figure 7.23 Distribution of the OA advantage by mean citation count

7.14. Self-citation counts

The rate of self-citation varies between OA and TA articles. Figure 7.24 shows by percentage article count the distribution of all categories of self-citation for both OA and TA articles. The data for OA articles has a mean of 3.43 and a standard deviation of 3.875 with a mode of 1 and a median of 2. For the TA data, the articles have a mean of 2.17 and a standard deviation of 2.746 with a modal value of 0 and a median of 1.

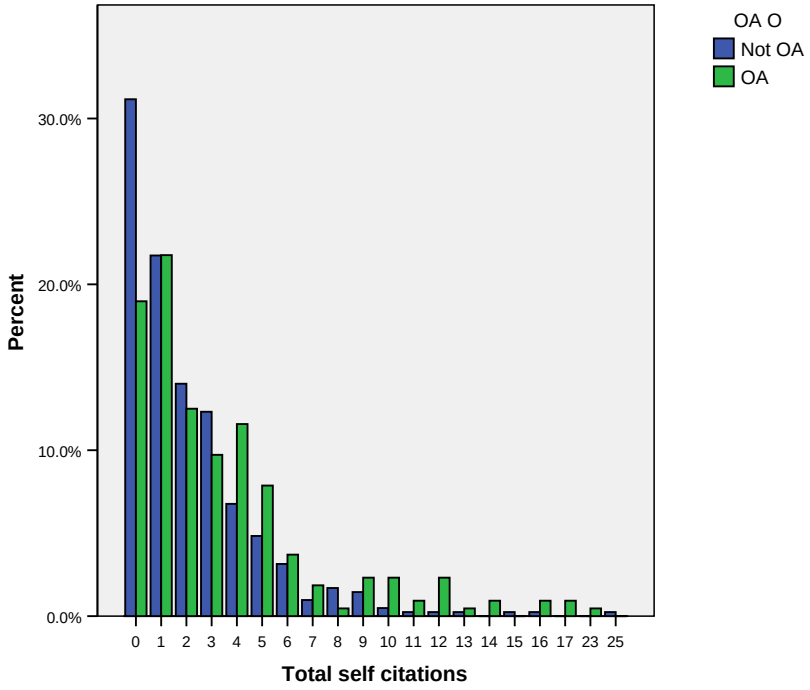


Figure 7.24 Distribution of self-citations by OA/TA status

Figure 7.25 shows the relative closeness of the self-citation counts when they are taken together. However, the most noticeable difference is apparent at the zero count, where 31.2% of TA articles have no self-citations, whereas for OA articles this is 19%.

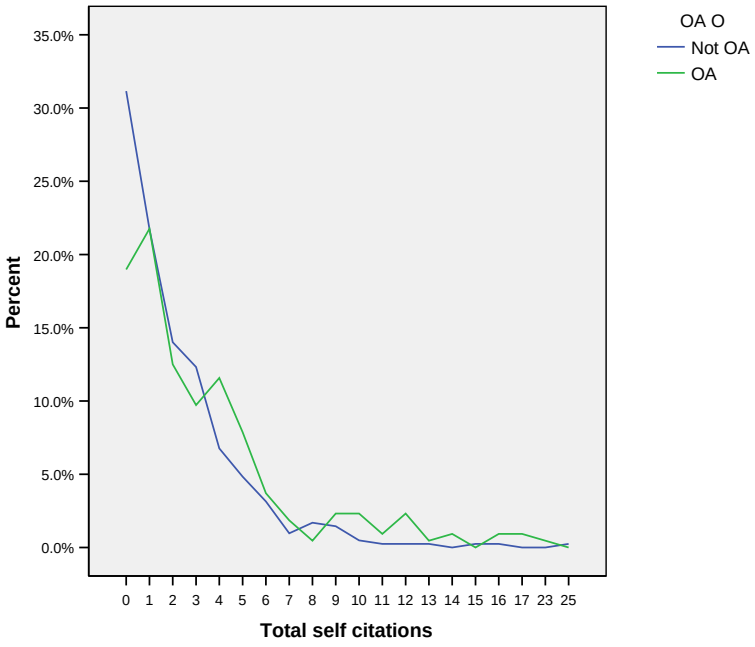


Figure 7.25 OA and TA citation distribution article

The boxplot in Figure 7.26 shows the distribution of self-citations by OA status. The median values for the self-citations are relatively close; the TA articles, however, have a marginally greater range.

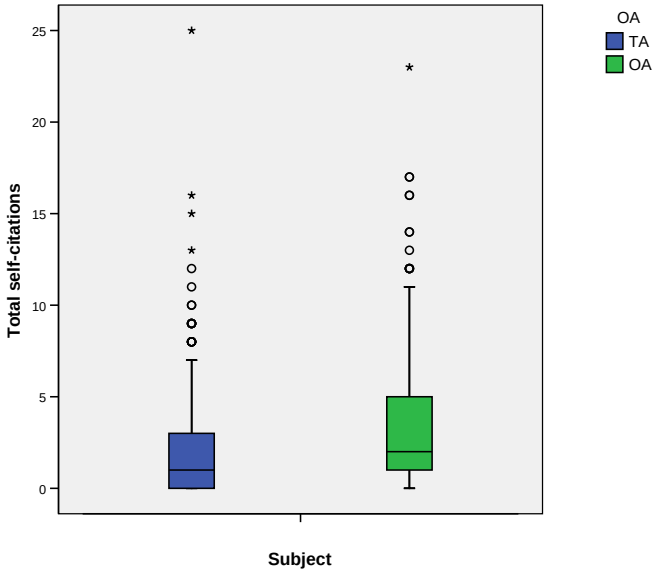


Figure 7.26 Boxplot of self-citations

Figure 7.27 show a breakdown of the gross citation count by the four types identified, three of which are related to author or journal self-citation.

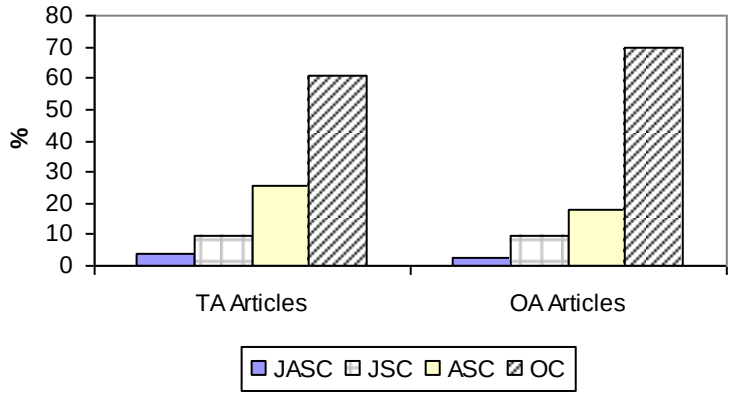


Figure 7.27 Breakdown of TA/OA citations

The combined self-citation rate for TA articles is 38.91% and for OA articles is 30.10%. Figure 7.28 shows a comparison between each citation category and their OA/TA status and illustrates that OA articles have a greater percentage of their citations in the 'other citation' category than TA articles. Conversely, TA articles have consistently more citations in the self-citation categories.

Figure 7.28 Articles by their citation category

However, OA articles consistently have a higher individual citation count for the articles within the self-citation categories outlined above, despite there being fewer OA articles

than TA articles in these self-citation categories. The mean number of journal and author self-citations for OA articles was 3.46, and for TA articles, this was 2.17. These means were compared using the independent 2 sample t-test; the result showed them to be from populations with different means ($p<0.001$).

The scatter plot shown in Figure 7.29 shows the relationship between other author citations (OC) and all types of self-citation. A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.29, the result was significant ($p<0.01$), for OA articles, this was 0.512 and for TA articles this was 0.490.

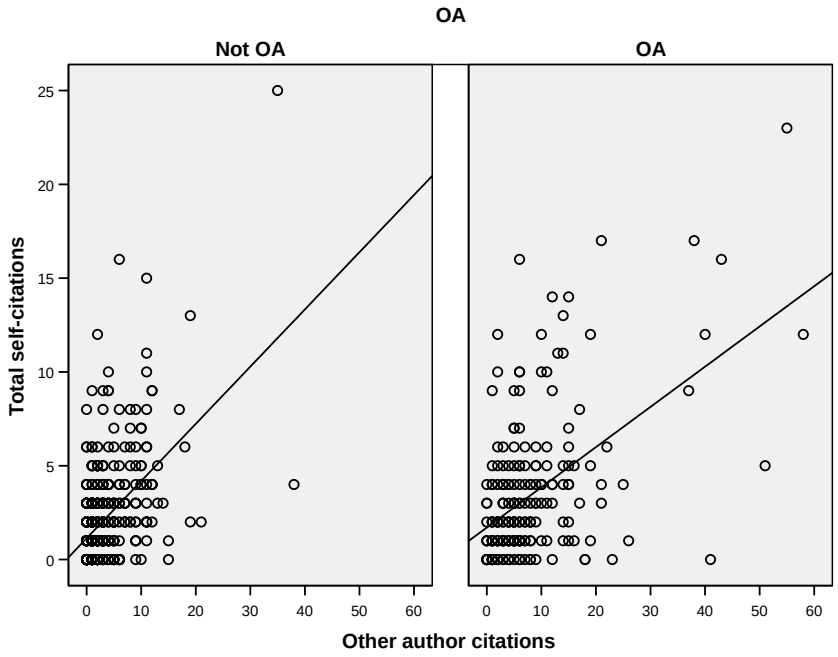


Figure 7.29 OA/TA scatterplot of self-citations to other author citations

7.15. Author frequency and OA/TA status

The mean number of authors for OA articles was 3.4 and for TA articles, this was 3.0. Figure 7.30 illustrates the distribution of author counts by the OA/TA status of their articles. The distribution is heavily skewed by the dominance of TA articles. In almost every case, there are more TA articles in each author count than OA articles. This is noticeably different to the sample taken in round one from high impact journals for this subject (ecology), which was almost evenly split by author counts between OA and TA articles.

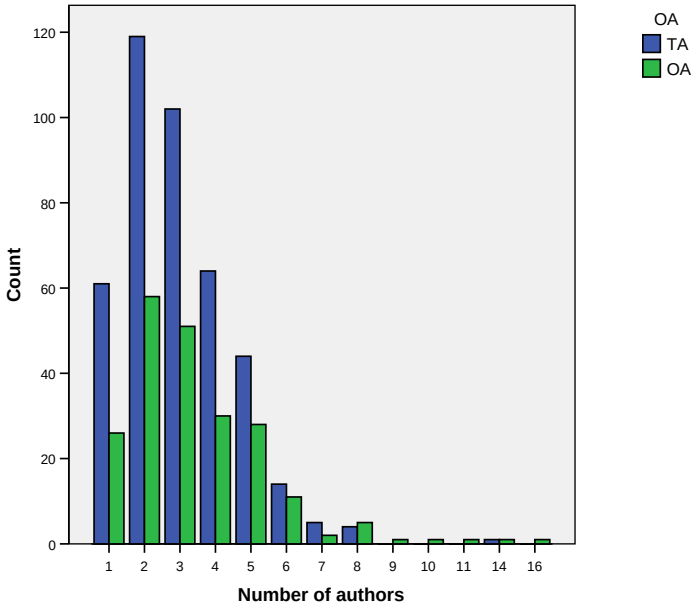


Figure 7.30 Articles by author count and OA/TA status.

The results from a Chi-square test showed there was no association overall between the number of authors and the OA/TA status of an article. The 630 articles sampled yielded by first author affiliation 58 countries of origin (two articles had no first author affiliation given). For analysis purposes, these were grouped into four regions: North America, continental Europe, UK and the Rest of the World.

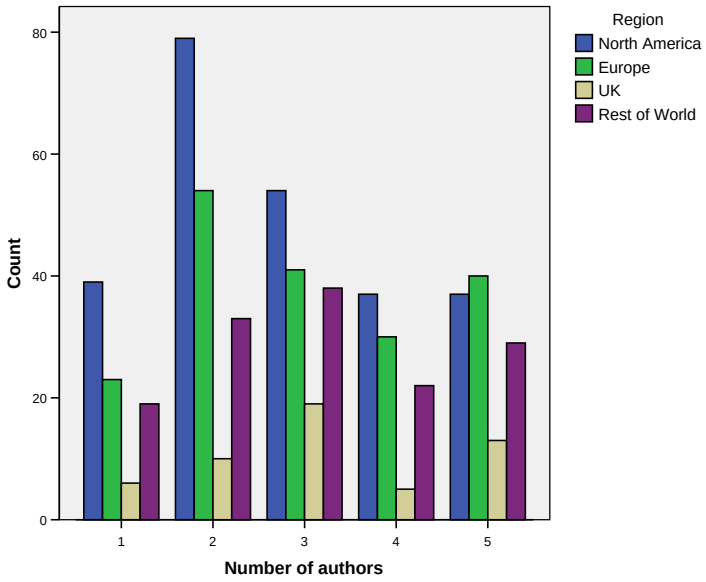


Figure 7.31 Articles by author count and region

Figure 7.31 shows the split when author counts are limited to no more than five, six and greater account for 3.5% of articles across all four regions. North America with 246

articles predominates followed by the Europe with 188 articles, the Rest of World with 141 and the UK with 53.

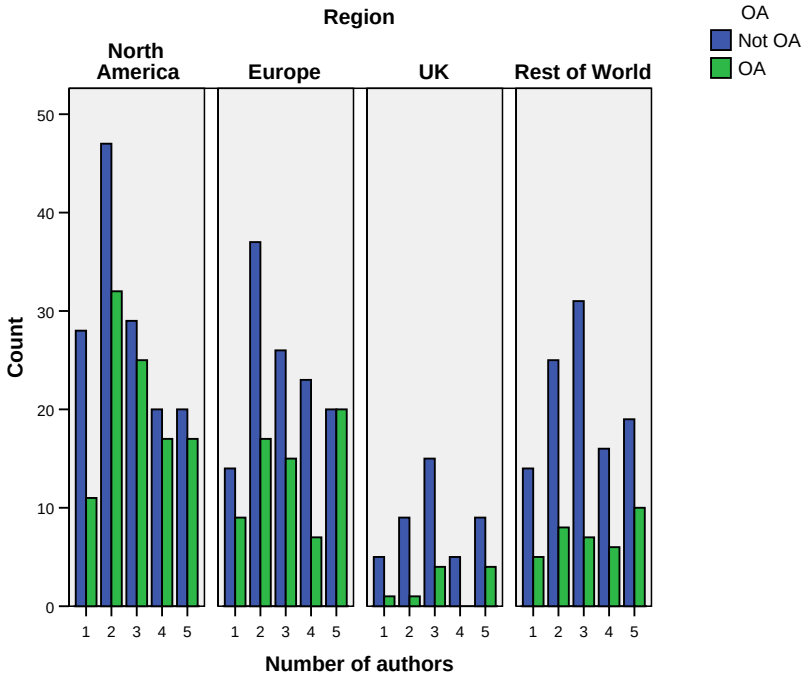


Figure 7.32 Author count by region and OA status

Figure 7.32 shows the frequency of TA articles in every region irrespective of author count and the dominance of the USA in the number of articles published. However, the distribution of articles between regions is less marked than in the purposive sample of ecology journal articles in the first round of data collection.

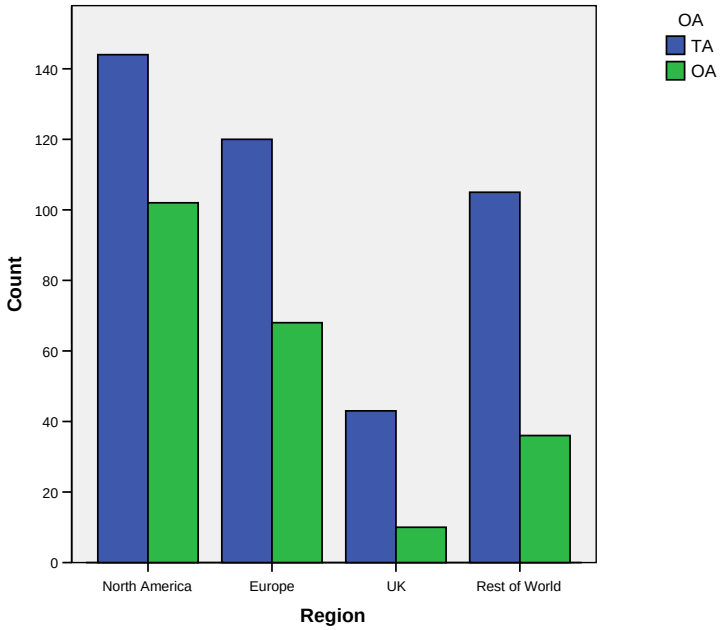


Figure 7.33 Number of OA/TA articles by region

Figure 7.33 collectively illustrates the dominance of TA articles with the UK and the Rest of the World having the highest percentage of TA articles at 81% and 74% respectively. North America has 59% of its articles TA and Europe has 64%.

7.16. Correlations

The pair of scatterplots in Figure 7.34 shows the distribution of citations against the number of authors for TA and OA articles.

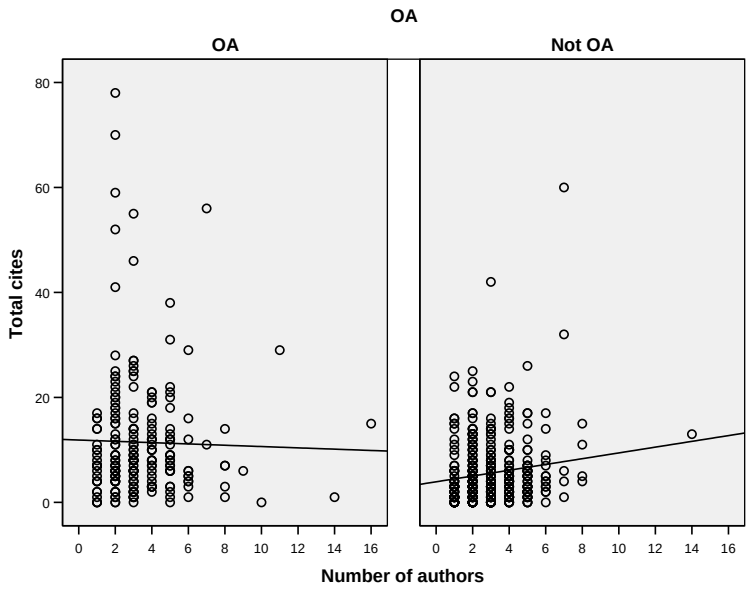


Figure 7.34 OA/TA author citation scatterplots

A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.34. For TA articles, this was 0.139 and for OA articles, this was negative at -0.022, however the OA result was not significant. Taking just journal and author self–citations and comparing this total to the level of authorship revealed no substantial differences between the two sets of data; the correlations were 0.198 and 0.133 for OA and TA articles respectively. Again, the OA result was not significant. Correlations between journal impact factor and the number of authors were also not significant. Perhaps not surprisingly, there is a significant correlation between impact factor and total citation counts. For OA articles, this was 0.382 and for TA articles this was 0.389. This correlation became marginally stronger for OA articles (0.409) and for TA articles (0.388) when just other author citations were taken into account.

7.17. Search engine success

Details of the how the search tools were used is given in 6.7. Figure 7.35 shows the relative success of *Google* and *Google Scholar* as compared to OAIster or OpenDOAR. Each of the bars in the chart represents the exclusive hits for that particular search tool.

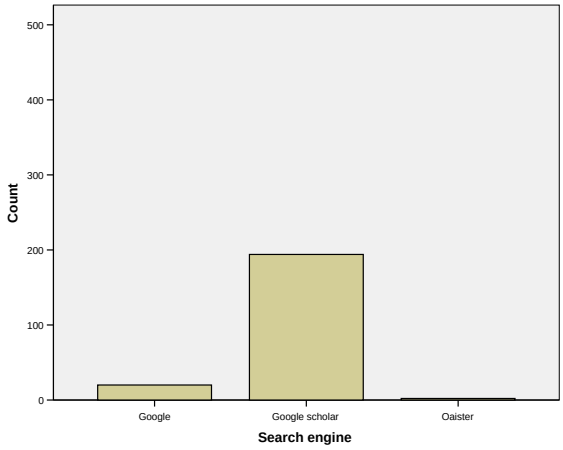


Figure 7.35 Search tool success rate

The percentage of records found for each search tool was; *Google* 9.30%, *Google Scholar* 89.80%, OAIster 0.90%, OpenDOAR found no records exclusively. The combined score for *Google* and *Google Scholar* was 99.10%. If compared to the first round results, although this second sample was randomly taken from all the journals listed in the *Journal Citation Reports* for 2004, there is a move away from the dominance of *Google* and *Google Scholar*. Table 6.9 shows the first round results by discovery tool and there is a

small difference in the share of finds for ecology, which was 80.42% for *Google Scholar* and 15.86% for *Google* with a combined score of 96.28%, the remaining 3.72% being shared between OAIster and OpenDOAR.

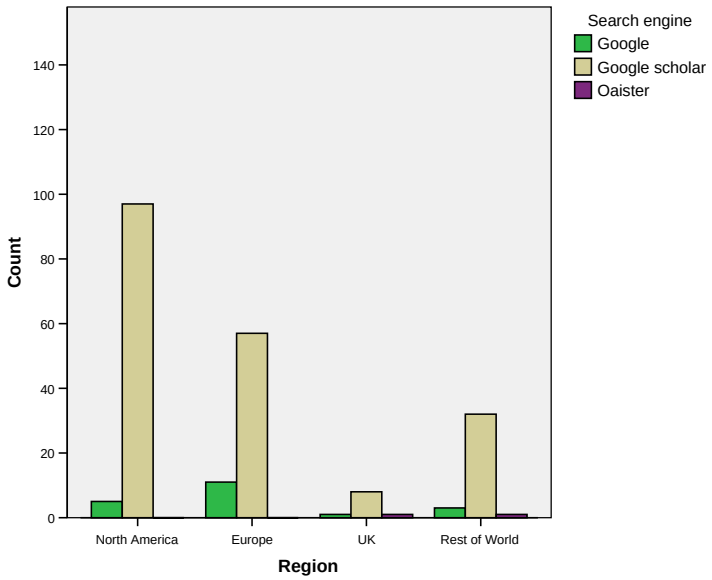


Figure 7.36 OA article hits by region and search tool

Taken at regional level, the hits in Figure 7.36 show the predominance of *Google Scholar* in North America.

		Region				
		North America	Europe	UK	Rest of World	Total
Google	Count	5	11	1	3	20
	% within Region	4.9%	16.2%	10.0%	8.3%	9.3%
Google scholar	Count	97	57	8	32	194
	% within Region	95.1%	83.8%	80.0%	88.9%	89.8%
Oaister	Count	0	0	1	1	2
	% within Region	.0%	.0%	10.0%	2.8%	.9%
Total	Count	102	68	10	36	216
	% within Region	100.0%	100.0%	100.0%	100.0%	100%

Table 7.4 Search tool success by subject and region

Figure 7.36 shows in detail the hits by percentage by search engine and territory. Whilst the counts shown in Figure 7.16 show the overall success of *Google Scholar*, Table 7.4

demonstrates just how small the regional differences are with the almost complete dominance of *Google* and *Google Scholar*.

7.18. *Impact factor*

Eighty percent of all impact factors for the 82 journals were in the range 0.191 to 3.804 (104 journals) with a mean of 2.43. There was only one extreme outlier 14.864 (*Trends in Ecology & Evolution*), which only accounted for five articles out of the 630 articles sampled. This is a far broader sample by impact factor than the purposive sample taken for ecology in the first round of data collection; see section 6.8.

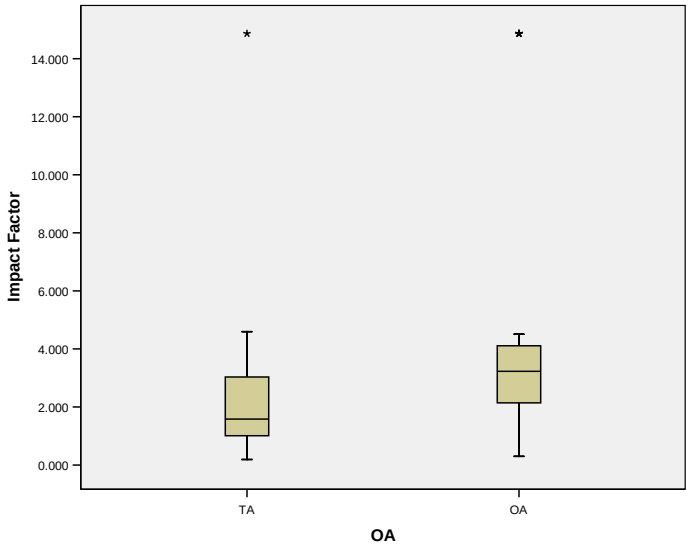


Figure 7.37 Impact factor by OA/TA status

The boxplot shown in Figure 7.37 indicates that the median and range of journal impact factors is higher for OA articles than TA articles, suggesting that OA articles are more likely to be found in higher impact factor journals. Asterisks indicate outliers that are more than three box-lengths away from the box, circles show outliers which are more than 1.5 box lengths away from the box.

The frequency of the OA/TA articles was plotted against their impact factor as shown in Figure 7.38 below. From the distribution can be seen a noticeable negative skew for the OA articles towards a greater frequency of OA articles appearing in the higher impact factor journals, confirming the result from Figure 7.37 above.

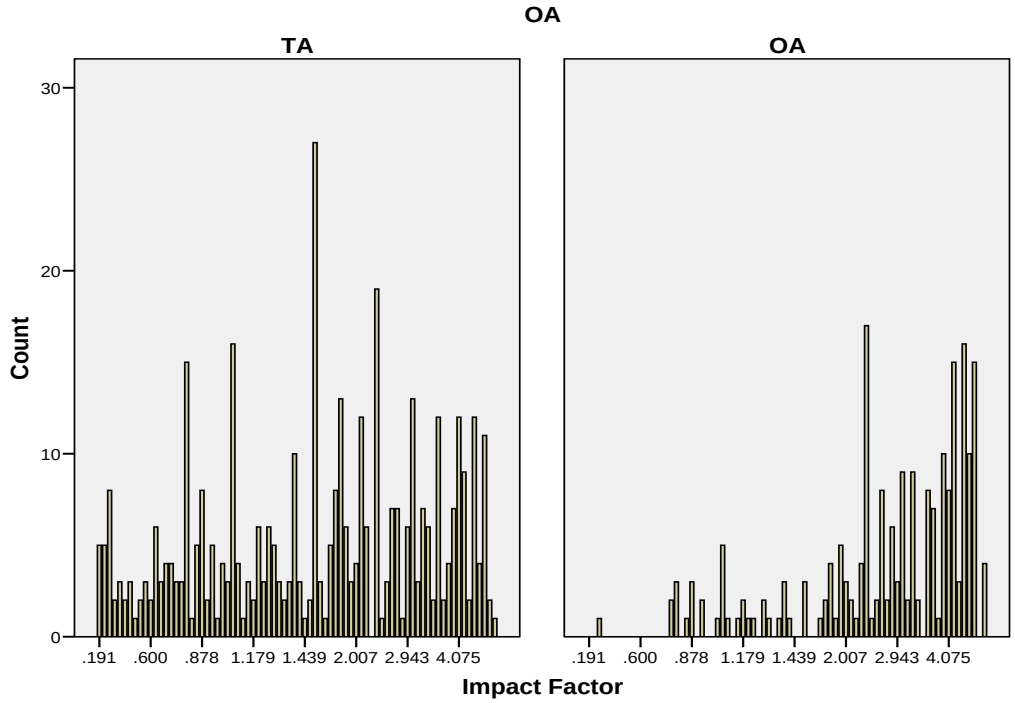


Figure 7.38 Impact factor by OA status

7.19. Within journal comparisons

Within journal comparisons for this random sample of articles for ecology differs from the earlier samples from the first round of data collection given the mean number of articles (7.68) from each of the 82 journals. However, there were similarities despite the randomness of the sample taken. Of the 82 journals sampled and their 630 articles, only 14 of the journals had more of their articles OA than TA. The distributions of citations within this subject and by OA/TA status is skewed in favour of a relatively small number of articles, which receive the majority of citations. Table 7.5 gives the overall distribution of citations by article count, OA/TA status, and also with and without self-citations.

Table 7.5 Distribution of citations by percentage article count

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All citations					
All articles	630	8.25	17.9	32.5	57.8
Toll Access	414	8.2	17.3	31.9	56.7
Open Access	216	8.33	19.44	36.1	64.4
Without self-citations					
All articles	630	6.0	14.4	27.5	52.5
Toll Access	414	7.0	14.3	26.5	50.0
Open Access	216	6.9	17.1	32.87	60.1

7.20. Citation distribution

Figure 7.39 shows the distribution of OA and TA ecology articles by their gross mean citation counts. The four smaller graphs show this by region. In all cases, there is a consistent difference in profile between OA and TA article citation counts. All show an advantage of OA over mean TA citation counts. There are some high impact journals amongst the random sample and those with an impact factor greater than four, some of which appeared in the first purposive sample were removed to give a view less skewed by high impact journals. The result shown in

Figure 7.40 shows, that the advantage is still consistently evident even when the high impact journals are removed. The graph for UK sample shown in

Figure 7.40 was, however, drawn from a small sample of OA articles and needs to be interpreted with caution.

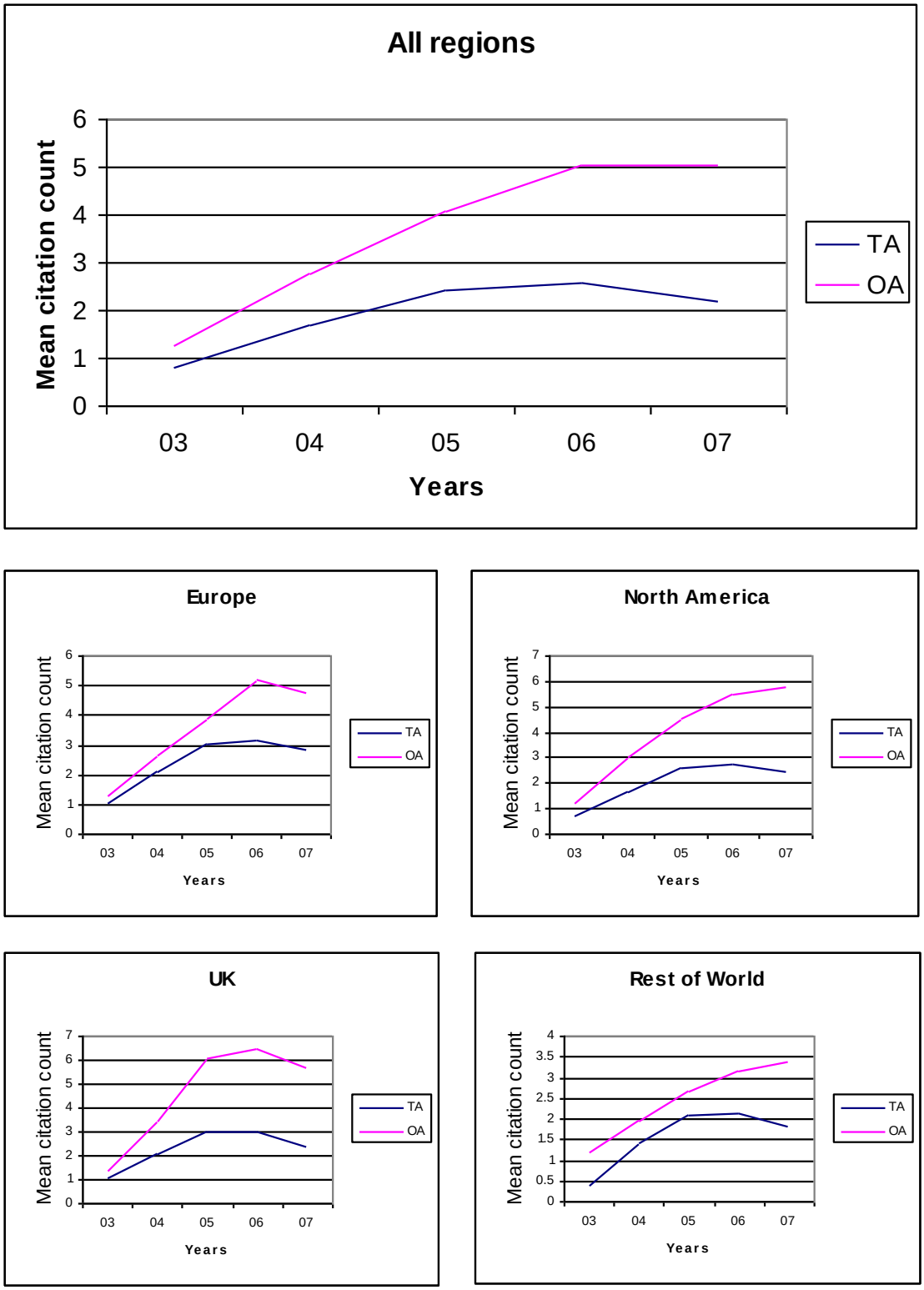


Figure 7.39 Gross mean citation profile by regions

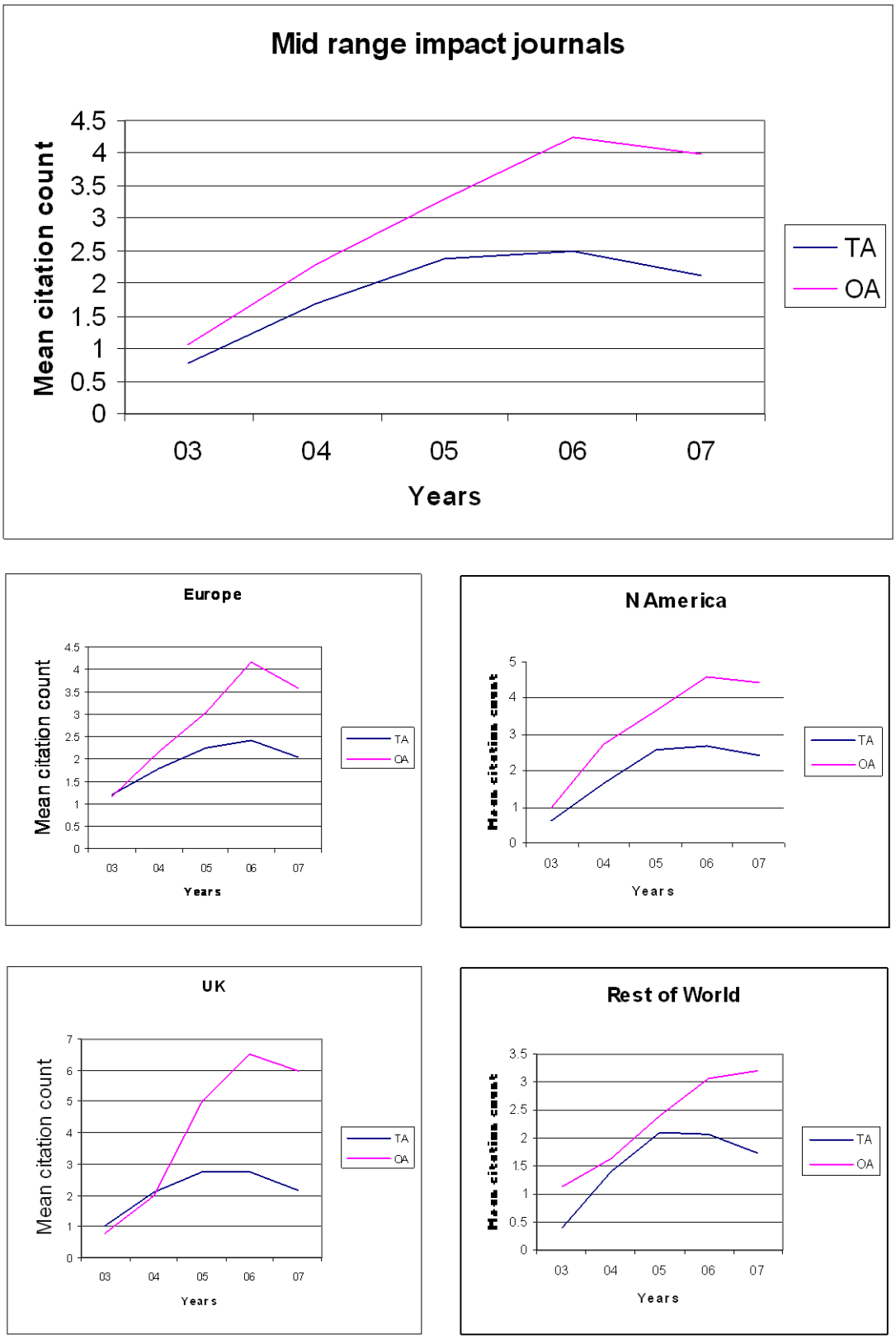


Figure 7.40 Lower impact journals by mean citation count

7.21. Objective 4

Objective 4 is concerned with determining if lower impact economics journals have the same OA/TA status and citation characteristics as their high impact counterparts.

7.22. Data overview

A purposive sample was taken from a range of TA journals around the mean impact factor for that discipline from the same dataset as that taken in the first round of data collection. The sample comprised of 980 article records from 21 journals. Of the 980 article, records 452 (46.1%) were TA and the remaining 528 (53.9%) were OA. The articles accrued 2634 citations; 274 articles did not receive any citations at all. Appendix F lists the journal titles from which the article records were taken.

7.23. Distribution of citation counts

Figure 7.41 illustrates the positive skew of citations when all of the 980 article citation records for both OA and TA records are counted and plotted. The data has a mean of 2.7 a standard deviation of 3.7, a mode of 0 and a median of 2. The citations counts are distributed such that 81.9% of all citations fall between 0 and 4. The overall range for citation counts is 0-48.

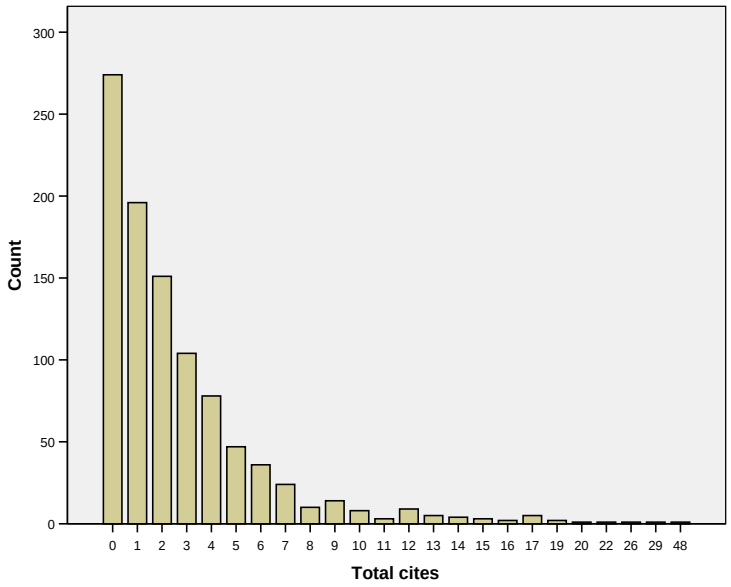


Figure 7.41 Distribution of all citations

The line graph at Figure 7.42 compares the citation counts for both the OA and TA articles. At the zero citation count, for TA articles 71.7% of citations fall between 0-2, whereas for OA articles, 56.3% of citations fall in this range. Of the 274 articles that did not attract any citations, the majority, 162 were TA (59%).

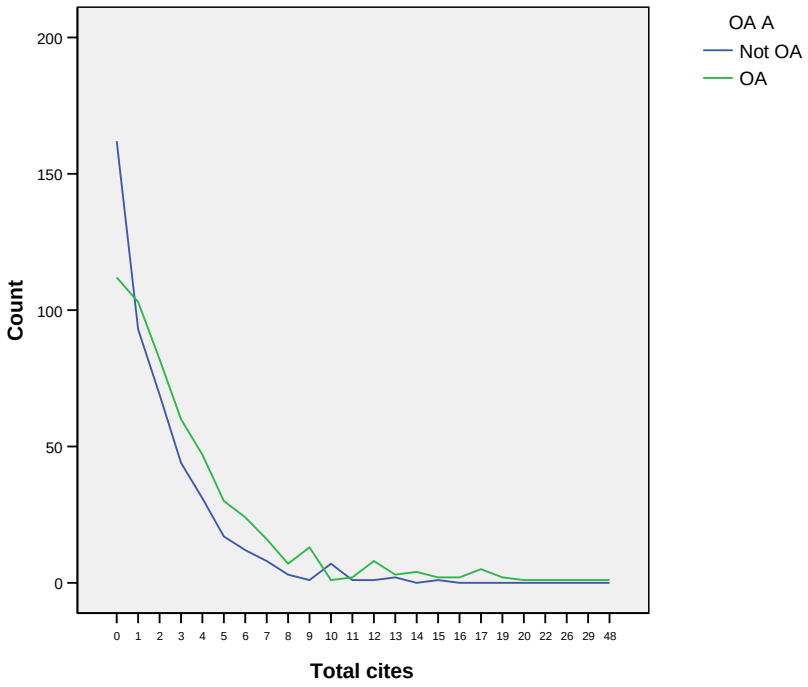


Figure 7.42 OA and TA citation distribution article

Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 3.35 compared to 1.92 for the TA articles. This gives a citation advantage in favour of OA articles of 74% (OA-TA/TA citation counts *100). The OA advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals. When these citations were excluded, the mean citation counts for the two article sets were OA 2.55 and TA 1.4. This extends the OA advantage, which becomes 82%.

A two sample independent *t*-test and Mann-Whitney test was carried out to test whether there was a statistically significant difference between the mean citation counts of OA and TA articles. In all cases, there was a significant difference between the means for both tests, indicating that the two populations of citation counts are drawn from two different populations ($p<0.001$).

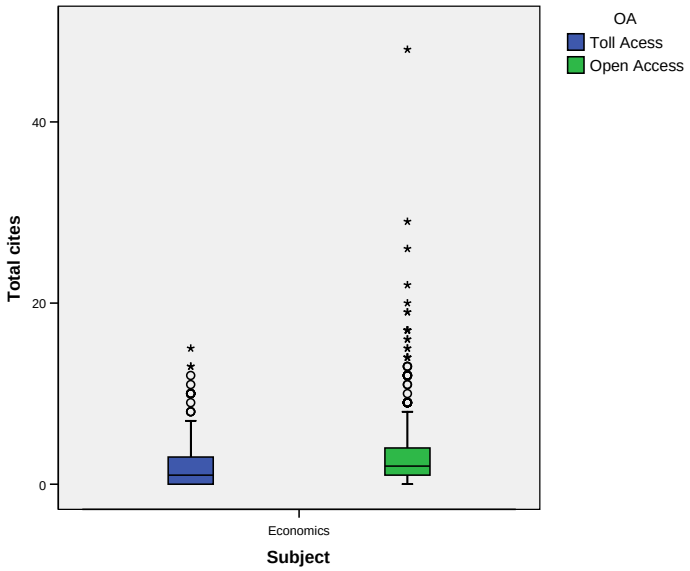


Figure 7.43 Boxplot of the distribution of all citations

The boxplot at Figure 7.43 helps illustrate the results from the *t*-test and the Mann-Whitney showing that OA articles have a marginally greater median citation value than TA articles. Asterisks indicate outliers that are more than three box-lengths away from the box, circles show outliers which are more than 1.5 box lengths away from the box. The boxplot in Figure 7.44 shows the distribution of other author citations by OA status.

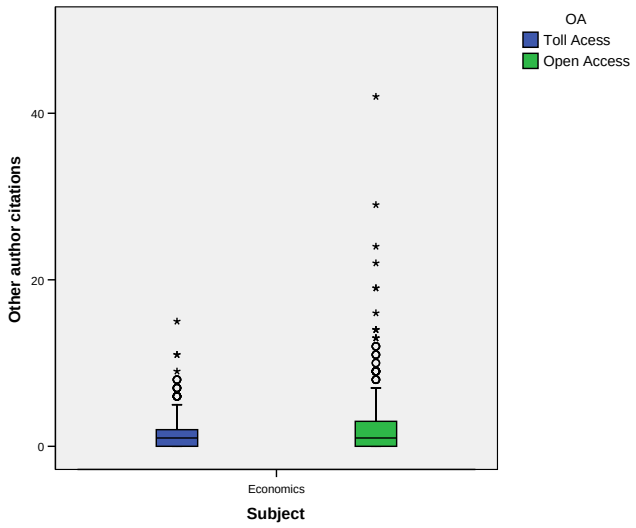


Figure 7.44 Boxplot of other author citations

7.24. Self-citation counts

The rate of self-citation varies between OA and TA articles. Figure 7.45 shows by percentage article count the distribution of all categories of self-citation for both OA and TA articles. The data for OA articles has a mean of 0.8 and a standard deviation of 1.3, with a mode of 0 and a median of 0. For the TA data, the articles have a mean of 0.52 and a standard deviation of 0.9, with a mode of 0 and a median of 0.

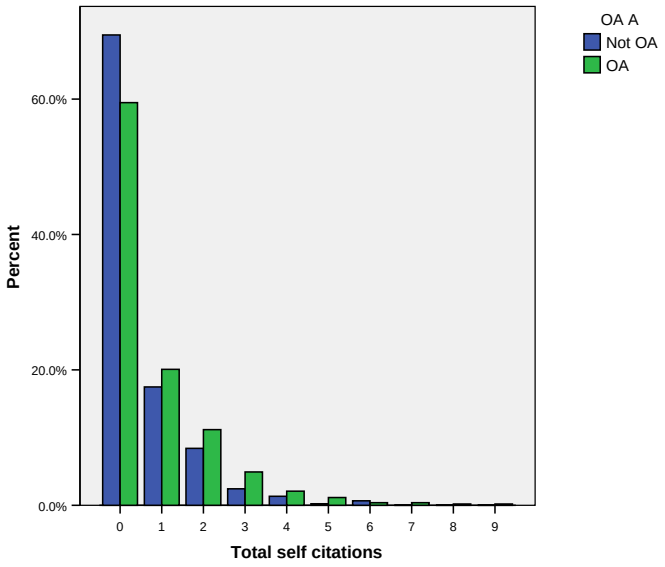


Figure 7.45 Distribution of self-citations by OA/TA status

Figure 7.46 shows the relative closeness of the self-citation counts when they are taken together. However, the most noticeable difference is apparent at the zero count, where 69.5% of TA articles have no self-citations, whereas for OA articles this is 59.5%.

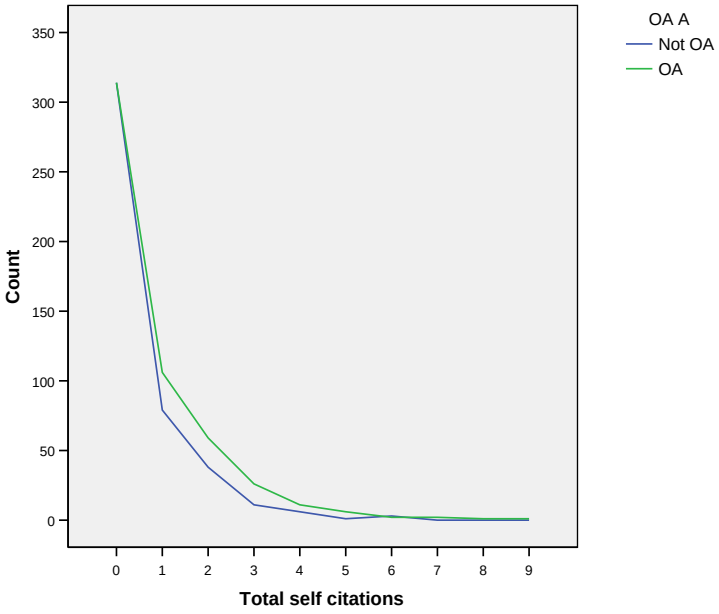


Figure 7.46 All OA and TA self citations

The boxplot in Figure 7.47 shows the distribution of self-citations by OA status. The median values for the self-citations are very close; the OA articles, however, have a greater range.

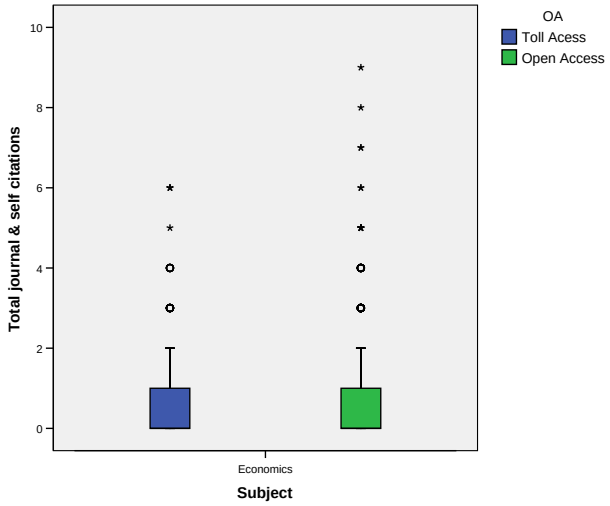


Figure 7.47 Boxplot of self-citations

Figure 7.48 show a breakdown of the gross citation count by the four types identified, three of which are related to author or journal self-citation.

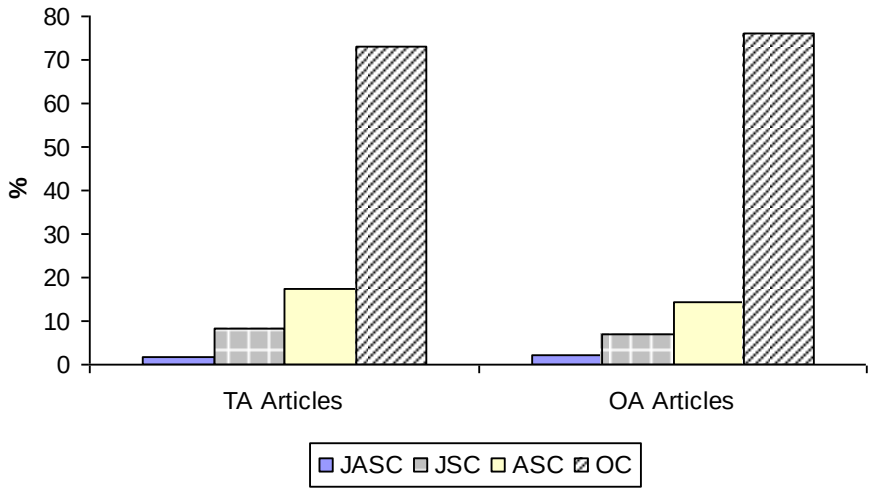


Figure 7.48 Breakdown of TA/OA citations

The combined self-citation rate for TA articles is 27% and for OA articles, this is 24%. Figure 7.49 shows a comparison between each citation category and their OA/TA status and illustrates that OA articles have a greater percentage of their citations in the 'other citation' category than TA articles. Conversely, TA articles have, apart from journal author self-citations (JASC), consistently more citations in the self-citation categories.

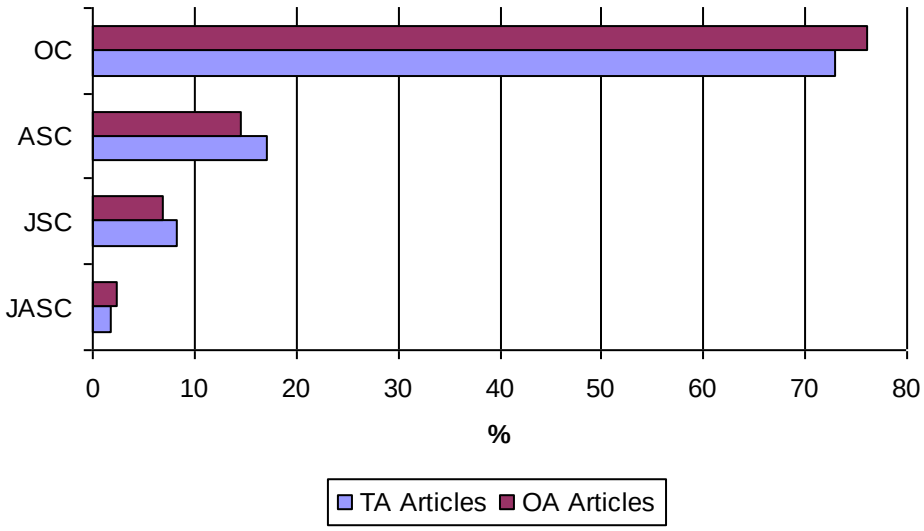


Figure 7.49 Articles by their citation category

However, OA articles consistently have a higher individual citation count for the articles within the self-citation categories outlined above, despite there being fewer OA articles than TA articles in these self-citation categories with the exception of journal author self-citations (JASC). The mean number of journal and author self-citations for OA articles was 0.79, and

for TA articles, this was 0.52. These means were compared using the independent 2 sample *t*-test; the result showed them to be from populations with different means ($p<0.001$).

The scatter plot shown in Figure 7.50 shows the relationship between other author citations (OC) and all types of self-citation. A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.50, the result was significant ($p<0.01$). For OA articles, this was 0.297 and for TA articles this was 0.235.

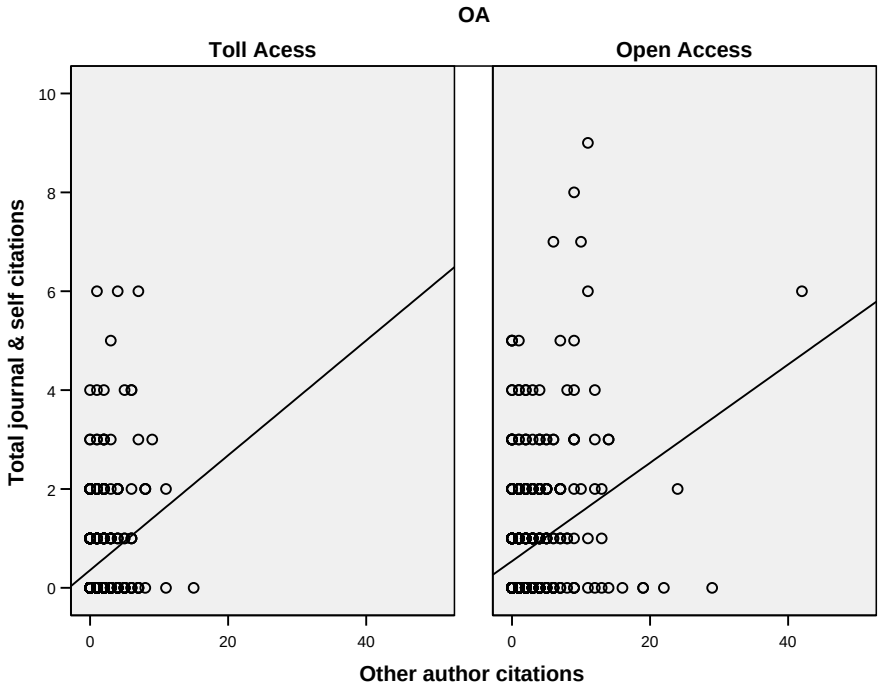


Figure 7.50 OA/TA scatterplot of self-citations to other author citations

7.25. Author frequency and OA/TA status

The mean number of authors for OA articles was 1.96 and for TA articles, this was 1.70. Figure 7.51 illustrates the distribution of author counts by the OA/TA status of their articles. There are noticeably more single-authored TA (225) articles than there are OA articles (178) despite there being more OA articles (528). Apart from this relatively small single author bias for TA articles, two, three and four author articles are more likely to be OA.

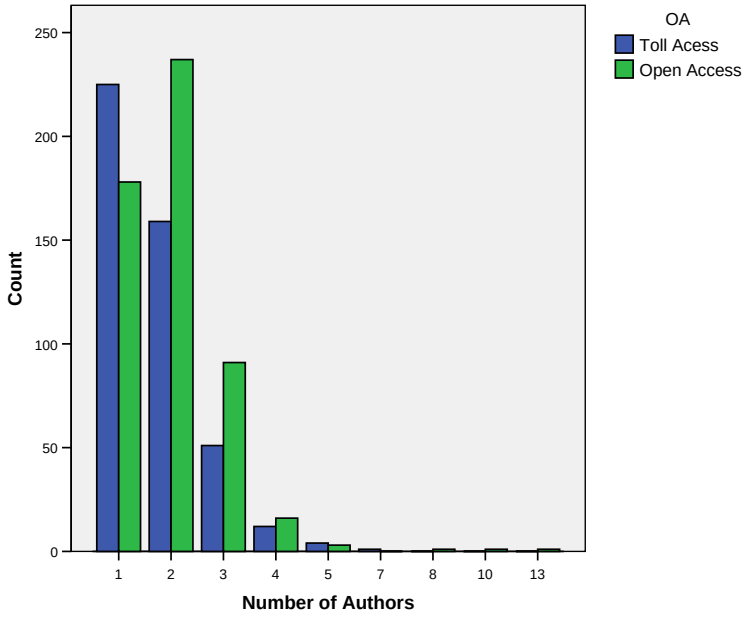


Figure 7.51 Articles by author count and OA/TA status.

The results of a Chi-square test ($\chi^2(4) =, 27.044, p<0.001$) showed there was a significant association overall between the number of authors and the OA/TA status of an article. However, the association between the number of authors and the OA status of an article showed that there was a tendency towards OA status only when there was either two or three authors. Hence, there is a strong association between single authorship and articles being TA. Of the 403 single authored articles 55.8% were TA. The 980 articles sampled yielded by first author affiliation 49 countries of origin (thirteen articles had no first author affiliation given). For analysis purposes, these were grouped into four regions: North America, continental Europe, UK and the Rest of the World.

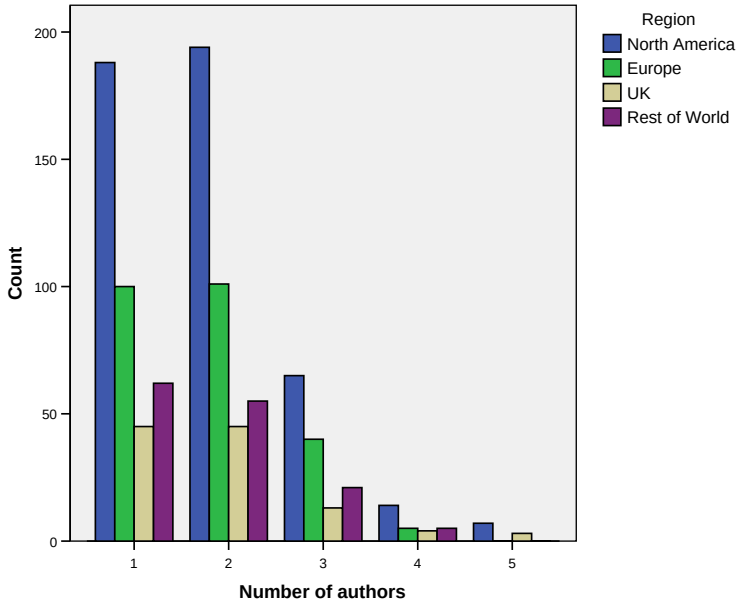


Figure 7.52 Articles by author count and region

Figure 7.52 shows the split when author counts are limited to no more than five, six and greater account for 0.4% of articles across all four regions. North America predominates with 468 articles followed by Europe with 246 articles, the Rest of the World with 143 and the UK with 110.

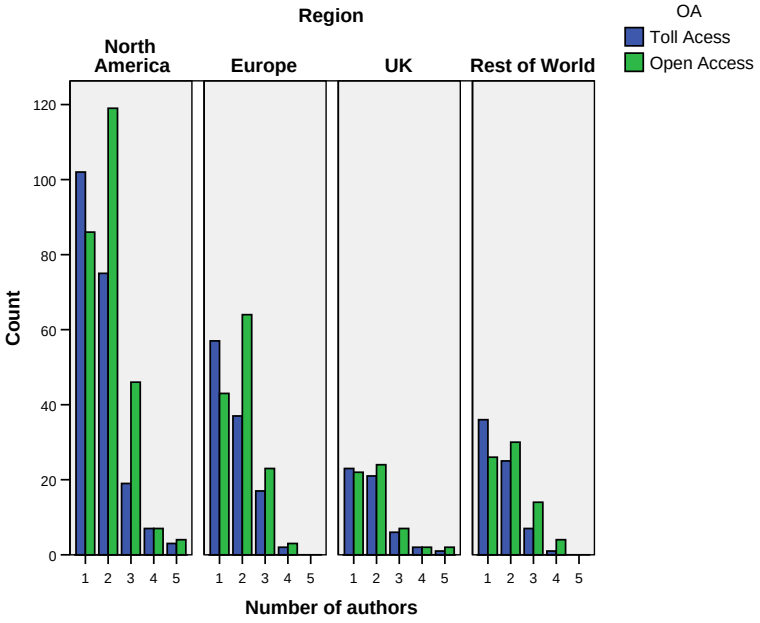


Figure 7.53 Author count by region and OA status

Figure 7.53 shows that there are a greater number of single authored TA articles in every region but thereafter multi-authored articles are OA and are either the same in number or greater than similarly authored TA articles.

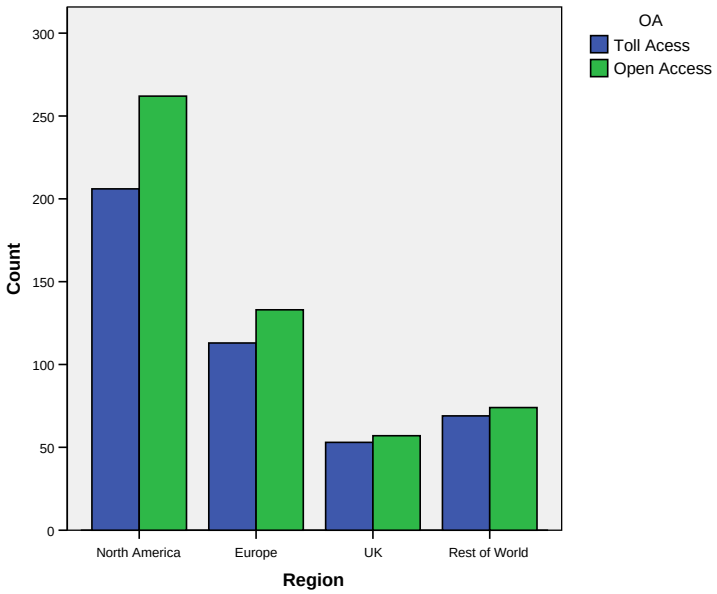


Figure 7.54 Number of OA/TA articles by region

Figure 7.54 collectively illustrates, however, the overall dominance of OA articles with North America and Europe having the highest percentage of OA articles at 56% and 54% respectively. Both the Rest of the World and the UK have 52% of its articles OA.

7.26. Correlations

The pair of scatterplots in Figure 7.55 shows the distribution of citations against the number of authors for TA and OA articles.

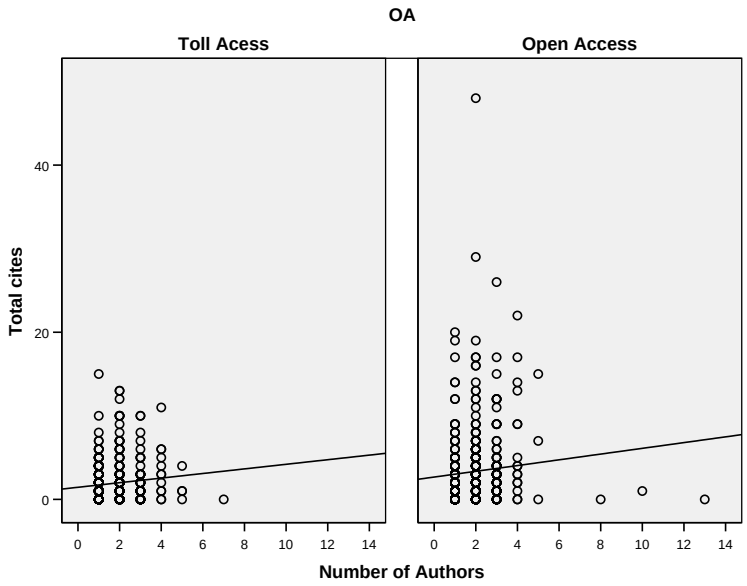


Figure 7.55 OA/TA author citation scatterplots

A one-tailed Pearson correlation coefficient was calculated for the variables shown in Figure 7.55. For TA articles, this was 0.099 and for OA articles, this was 0.081 ($p<0.05$). Taking just journal and author self–citations and comparing this total to the level of authorship revealed a poor correlation; the correlations were 0.034 and 0.069 for TA and OA articles respectively. The results were not significant. The correlations between journal impact factors and the number of authors was also not significant. Likewise, the correlation between impact factor and total citation counts was weak for TA articles at 0.111 and for OA articles it was very weak at 0.016 and not significant.

7.27. Search engine success

Details of the how the search tools were used is given in section 6.7. Figure 7.15 shows the relative success of *Google* and *Google Scholar* as compared to OAIster or OpenDOAR. Apart from the combined OAIster and OpenDOAR bar, each of the bars in the chart represents the exclusive hits for that particular search tool. For the combined OAIster and OpenDOAR entry, this is where both of them located the same OA articles.

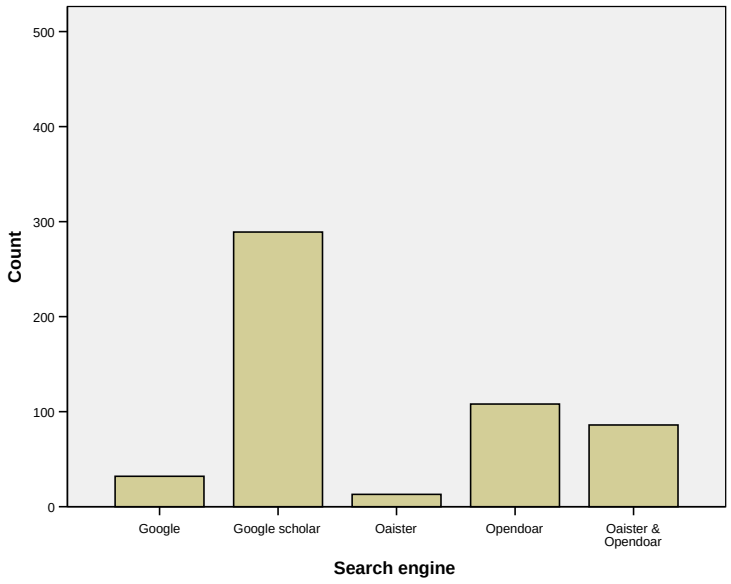


Figure 7.56 Search tool success rate

The percentage of records found for each search tool was; *Google* 6.1%, *Google Scholar* 54.7%, OAIster 2.5%, OpenDOAR 20.5% and where OAIster and OpenDOAR retrieved the same article, their combined score was 16.2%. The combined score for *Google* and *Google Scholar* was 60.8%. If compared to the first round results, although not an absolute like for like match, there is a move away from the dominance of *Google* and *Google Scholar*. Table 6.9 shows the first round results by discovery tool and there is a notable difference in the score for economics, which was 71.9% for *Google Scholar* and 6.9% for *Google* with a combined score of 78.8%. The combined score for OAIster and OpenDOAR was 21.2%. Comparing this to the second round shows a move of 17.9% points towards finding being able to find these OA articles in either OAIster or Open DOAR.

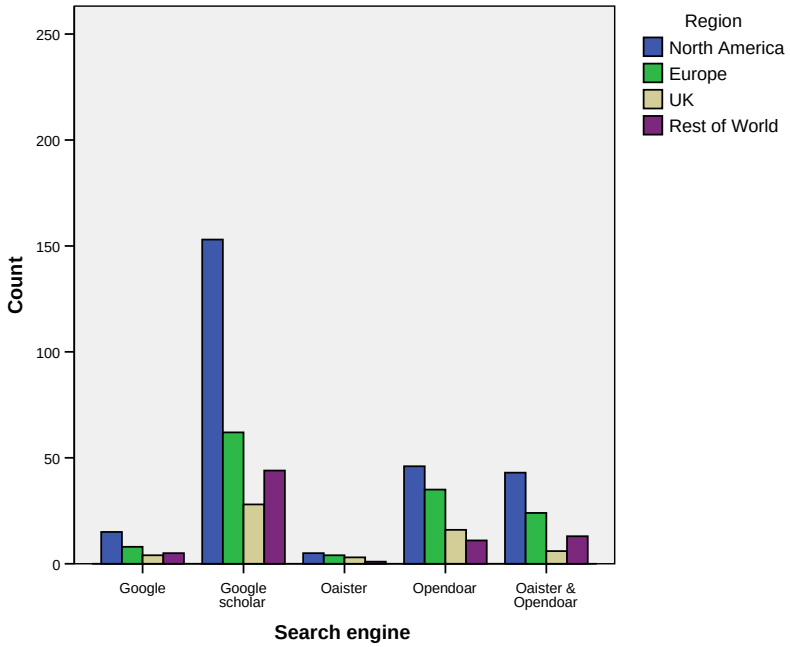


Figure 7.57 OA article hits by region and search tool

Taken at regional level, the hits in Figure 7.57 shows the predominance of *Google Scholar* in North America. Clearly however, most of the OA article hits were located by first author affiliation in that region. More noticeable in this subject (economics) is the relative success of OAIster and OpenDOAR.

		Region				
		North America	Europe	UK	Rest of World	Total
Google	Count	15	8	4	5	32
	% within Region	5.7%	6.0%	7.0%	6.8%	6.1%
Google scholar	Count	153	62	28	44	287
	% within Region	58.4%	46.6%	49.1%	59.5%	54.6%
Oaister	Count	5	4	3	1	13
	% within Region	1.9%	3.0%	5.3%	1.4%	2.5%
Opendoar	Count	46	35	16	11	108
	% within Region	17.6%	26.3%	28.1%	14.9%	20.5%
Oaister & Opendoar	Count	43	24	6	13	86
	% within Region	16.4%	18.0%	10.5%	17.6%	16.3%
Total	Count	262	133	57	74	526
	% within Region	100.0%	100.0%	100.0%	100.0%	100.0%

Table 7.6 Search tool success by subject and region

Figure 7.57 shows in detail the hits by percentage by search engine and territory. Whilst the counts shown in Figure 7.57 show the overall success of *Google Scholar*, Table 7.6

demonstrates the regional differences and the continued, and increasing success from the first round data for OAIster and OpenDOAR.

7.28. Impact factor

The 21 journals had impact factors ranging from 0.398 to 0.967; this is noticeably narrower than the impact factors for the journals from the first round of data collection for this subject (economics), which ranged from 1.345 to 3.222. The boxplot shown in Figure 7.58 indicates that the range of journal impact factors for OA and TA articles are very closely matched, suggesting for this sample of mid-impact journal titles that OA articles are not associated with mid impact journals. Asterisks indicate outliers that are more than three box-lengths away from the box, and circles show outliers which are more than 1.5 box lengths away from the box. A Chi-squared test on the association of impact factor and OA status of an article ($\chi^2(16) = 56.132$, $p < .001$) confirms that neither OA nor TA articles are more associated with these mid impact journals.

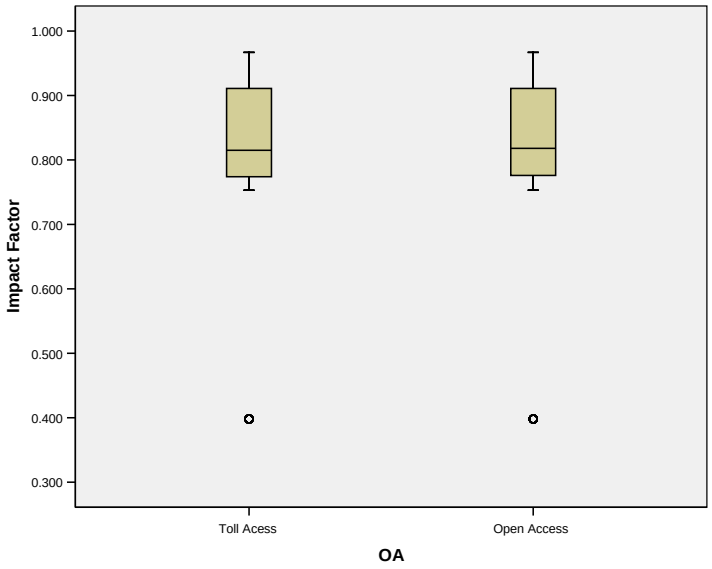


Figure 7.58 Impact factor by OA/TA status

7.29. Within journal comparisons

Figure 7.59 illustrates the split between OA and TA articles within the journals for economics.

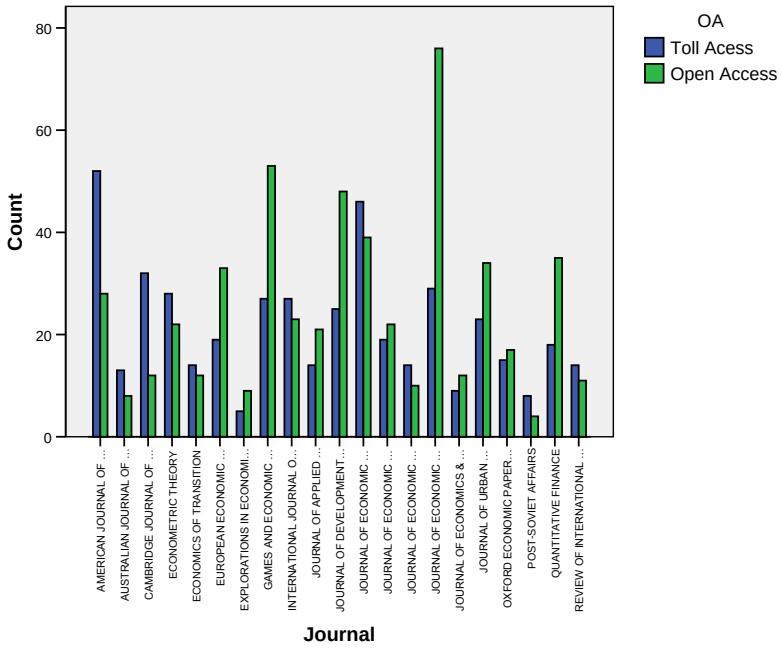


Figure 7.59 OA/TA article split for economics

The split for economics is roughly even between OA and TA journals, with 11 of the 21 journals having more OA articles than TA articles.

7.30. Distribution of citations within subjects

The distributions of citations within this subject and by OA/TA status is skewed in favour of a relatively small number of articles, which receive the majority of citations. Table 7.7 gives the overall distribution of citations by article count, OA/TA status, and also with and without self-citations.

Table 7.7 Distribution of citations by percentage article count

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All citations					
All articles	980	5.9	14.1	26.7	48.6
Toll Access	452	6.6	14.6	26.1	44.9
Open Access	528	6.0	14.7	28.4	52.3
Without self-citations					
All articles	980	5.0	11.9	23.1	41.9
Toll Access	452	5.7	12.3	23.2	40.2
Open Access	528	5.3	12.4	24.2	45.1

Taking the article records from the *Journal of Economic Behavior & Organisation* and splitting these by their OA/TA status shows that the characteristics exhibited in Figure 7.45 occur at single journal level as well. Table 7.8 shows the distribution of citations for the *Journal of Economic Behavior & Organisation* both with and without self-citations. At 85 articles, this is a relatively small sample.

Table 7.8 Distribution of citations by percentage article count at journal level

Citation category	Total article count	% of articles accounting for 30% of citations	% of articles accounting for 50% of citations	% of articles accounting for 70% of citations	% of articles accounting for 90% of citations
All citations					
All articles	85	7.5	15.3	27.1	45.8
Toll Access	46	8.2	15.6	28.2	45.1
Open Access	39	7.1	14.5	23.9	41.0
Without self-citations					
All articles	85	6.5	12.9	22.4	37.6
Toll Access	46	8.0	15.3	26.0	41.3
Open Access	39	7.7	15.4	23.1	35.9

7.31. Objective 5.

This objective is concerned with trying to determine if causal links can be found between any OA citation advantage and, for example, the bibliographic details of the citing items. The article records collected for applied mathematics from the first round of data collection were extended by recording the bibliographic details of those authors that cite them and their first author affiliation by country. These extended article records and some of the other first round data was examined using statistical techniques already employed and, in some cases, using logistic regression.

Logistic regression is a statistical technique like multiple regression but which has an outcome variable that is "a categorical dichotomy and predictor variables that are continuous or categorical" (Field 2005, p.218). In the case of journal articles here, the categorical dichotomy is whether they are OA or TA and the predictor variables, which might help predict the OA status of an article could be, for example, the number of authors, the subject, or the country of origin of the author(s). The technique helps determine which, if any of the predictor variables have an influence on the OA status of an article. Usefully, logistic regression makes no assumptions about the distribution of predictor variables: "the predictors do not have to be normally distributed, linearly related or of equal variance within each group" (Tabachnick & Fidell 2001, p.517). The technique was employed by Eysenbach (2006) in his important analysis of OA articles published in *Proceedings of the National Academy of Sciences*.

In the first round of data collection, 1158 article records were identified from high impact journals for applied mathematics during September 2006. These articles collected, excluding self-citations and journal self-citations, 2919 other author citations. This data was extended in mid 2007 by collecting the bibliographic details of citing authors. The citation count, to the end of 2006, had increased the other author citation count to 3032 (i.e. omitting self-citations). These 3032 citation records were linked to each of their respective articles from the original 1158 articles identified.

7.32. Countries of origin

Of the original 1158 articles collected for applied maths, 480 were TA and the remaining 678 were OA. After all the non-cited articles and those which contained only author and journal self-citations were removed, this left 793 articles of which 298 were TA and 495 were OA. Table 7.9 shows, by first author affiliation, the 793 cited articles by their region of origin. The USA, European countries and the rest of Europe made up 84.9% of all the articles.

Table 7.9 Cited articles by region and OA status

			OA		
			Toll Access	Open Access	Total
	Spain	Count	12	14	26
		% within Region	46.2%	53.8%	100.0%
	Japan	Count	16	9	25
		% within Region	64.0%	36.0%	100.0%
	Italy	Count	20	22	42
		% within Region	47.6%	52.4%	100.0%
	Germany	Count	18	47	65
		% within Region	27.7%	72.3%	100.0%
	France	Count	26	39	65
		% within Region	40.0%	60.0%	100.0%
	Canada	Count	11	16	27
		% within Region	40.7%	59.3%	100.0%
	Pacific Rim	Count	9	22	31
		% within Region	29.0%	71.0%	100.0%
	China	Count	15	11	26
		% within Region	57.7%	42.3%	100.0%
	Rest of World	Count	20	18	38
		% within Region	52.6%	47.4%	100.0%
	UK	Count	18	28	46
		% within Region	39.1%	60.9%	100.0%
	Europe	Count	31	65	96
		% within Region	32.3%	67.7%	100.0%
	USA	Count	102	204	306
		% within Region	33.3%	66.7%	100.0%
Total		Count	298	495	793
		% within Region	37.6%	62.4%	100.0%

Table 7.10 shows the corresponding table to Table 7.9, here is shown how the 3032 other author citations to the cited articles are broken down by both their OA status and the region from which the first affiliated author came from.

Table 7.10 Citations to cited articles by region and OA status

			OA		
			Toll Access	Open Access	Total
	Spain	Count	39	63	102
		% within region	38.2%	61.8%	100.0%
	Japan	Count	52	90	142
		% within region	36.6%	63.4%	100.0%
	Italy	Count	81	120	201
		% within region	40.3%	59.7%	100.0%
	Germany	Count	70	214	284
		% within region	24.6%	75.4%	100.0%
	France	Count	59	137	196
		% within region	30.1%	69.9%	100.0%
	Canada	Count	27	60	87
		% within region	31.0%	69.0%	100.0%
	Pacific Rim	Count	51	124	175
		% within region	29.1%	70.9%	100.0%
	China	Count	90	140	230
		% within region	39.1%	60.9%	100.0%
	Rest of World	Count	55	104	159
		% within region	34.6%	65.4%	100.0%
	UK	Count	60	114	174
		% within region	34.5%	65.5%	100.0%
	Rest of Europe	Count	120	321	441
		% within region	27.2%	72.8%	100.0%
	USA	Count	231	610	841
		% within region	27.5%	72.5%	100.0%
Total		Count	935	2097	3032
		% within region	30.8%	69.2%	100.0%

There is an overall shift of almost seven percentage points in terms of the number of citing articles, inasmuch that 69.2% of these articles cite the 62.4% of cited OA articles. This shift is indicated by the varying differences in percentage citation counts and most notably in Japan, China and the Rest of the World, which have a greater percentage share of the citing articles by first author affiliation than their respective cited article counts. The OA advantage is maintained overall at 35.0% although this is somewhat reduced from the OA advantage of 53% shown in Table 6.1.

In some cases, there are multiple citations to the cited articles from some of the citing articles. Table 7.11 shows how the 3032 citations from the 2680 citing articles that cited the

original 793 were broken down. For example, there were 2413 citing articles that cited just one of the 793 articles, whereas there were three articles, which cited six of the original articles each. First author affiliations for the 793 articles covered 47 countries. The first author affiliation of the 2680 citing articles, citing the 793 articles were drawn from 70 countries; 23 of these were evidently new countries. Initially the cited and citing countries were classified by their per capita income using The World Bank's (2007) system of classification. For example, China is designated as being in the lower middle-income group of countries and India in the low-income bracket, whereas most of Western Europe and North America are in the high-income group of countries. To further aid analysis, the original 47 countries and the 70 citing author countries were classified by location into USA, Canada, France, Germany, Italy, Japan, Spain, UK, rest of Continental Europe, China, Pacific Rim, and the Rest of World.

Table 7.11 Frequency of citation

Frequency of Citation	Citing Articles	Overall Citations
1	2413	2413
2	208	416
3	43	129
4	9	36
5	4	20
6	3	18
Totals	2680	3032

Table 7.12 shows the distribution of the original 793 cited articles and the distribution of the 3032 citations by the country of their first author affiliation. Countries are classified by their World Bank per capita income grouping. The number of articles appearing in each category has been given by its occurrence and the ratio is the division of the citing articles by the cited articles. The numbers in brackets indicate the number of article records in each category.

Table 7.12 Ratio of citations to cited articles by income group

Status	**World Bank classification by per capita income**			
TA Articles	Low	Lower middle	Upper middle	High
TA cited articles (298)	2	21	19	256
TA citing articles (935)	6	106	75	748
Ratio of citing to cited articles	3	5.1	3.9	2.9
OA Articles				
OA cited articles (495)	1	14	18	462
OA citing articles (2097)	17	180	126	1774
Ratio of citing to cited articles	17.0	12.9	7.0	3.8

The results suggest a fairly stable ratio between the TA cited and citing articles but a more noticeable variation for OA articles where there is a higher ratio of articles which cite OA articles in the lower income groupings.

Table 7.13 shows just the distribution of citations to cited articles by their OA status and per capita income group. What is evident is that there is a greater percentage of citations to the TA articles (20.00%) from the low to upper middle income groups than is the case for OA articles where the comparable group of is 15.4% of the 2097 citations to the OA articles.

Table 7.13 Citations to citing articles by income group

OA				Cited Country Income				Total
				Low	Lower middle	Upper middle	High	
Toll Access	Citing Country Income	Low	Count	0	1	0	5	6
			% within Citing Country Income	.0%	16.7%	.0%	83.3%	100.0%
		Lower middle	Count	3	22	8	73	106
			% within Citing Country Income	2.8%	20.8%	7.5%	68.9%	100.0%
		Upper middle	Count	0	6	15	54	75
			% within Citing Country Income	.0%	8.0%	20.0%	72.0%	100.0%
		High	Count	2	50	19	677	748
			% within Citing Country Income	.3%	6.7%	2.5%	90.5%	100.0%
	Total		Count	5	79	42	809	935
			% within Citing Country Income	.5%	8.4%	4.5%	86.5%	100.0%
Open Access	Citing Country Income	Low	Count	0	1	0	16	17
			% within Citing Country Income	.0%	5.9%	.0%	94.1%	100.0%
		Lower middle	Count	0	10	5	165	180
			% within Citing Country Income	.0%	5.6%	2.8%	91.7%	100.0%
		Upper middle	Count	1	3	6	116	126
			% within Citing Country Income	.8%	2.4%	4.8%	92.1%	100.0%
		High	Count	1	44	27	1702	1774
			% within Citing Country Income	.1%	2.5%	1.5%	95.9%	100.0%
	Total		Count	2	58	38	1999	2097
			% within Citing Country Income	.1%	2.8%	1.8%	95.3%	100.0%

The 3032 citations to the 793 articles are shown in Table 7.14. They are matched by the country of first author affiliation. For example, of the 231 TA citations by authors affiliated to American institutions, just under half (115) were from other American affiliated authors. By contrast, of the 140 OA citations by authors affiliated to Chinese institutions, only seven were from other Chinese affiliated authors. Overall, only 26.0% of TA articles had matching cited and citing records by first author affiliation. For OA articles, this percentage was 27.1.

Table 7.14 Regional citation match by author country

Count

OA		Region to region match: no match	Region to region match: match	Total
Toll Access	USA	116	115	231
	Rest of Europe	104	16	120
	UK	48	12	60
	Rest of World	50	5	55
	China	70	20	90
	Pacific Rim	44	7	51
	Canada	26	1	27
	France	45	14	59
	Germany	59	11	70
	Italy	65	16	81
	Japan	28	24	52
	Spain	37	2	39
Total		692	243	935
Open Access	USA	251	359	610
	Rest of Europe	266	55	321
	UK	97	17	114
	Rest of World	99	5	104
	China	133	7	140
	Pacific Rim	114	10	124
	Canada	55	5	60
	France	108	29	137
	Germany	168	46	214
	Italy	93	27	120
	Japan	84	6	90
	Spain	60	3	63
Total		1528	569	2097

Table 7.15 and Table 7.16 show similar results to that of Table 7.14, except the cross-tabulation matches, in detail, cited to citing records by country or region such that the distribution of all citations is shown. For example, there are 344 citations from the other eleven countries or regions and the USA itself to TA articles written by first author affiliation from the USA. Collective citation ratios for OA to TA articles for each region or country are generally about two and above for those countries from Western Europe. For the USA, the ratio is 2.7 in favour of OA articles (928OA/344TA articles). The OA/TA citation ratio for the Rest of World and for China is around one, and approaching two for the Pacific Rim countries. A notable exception is Japan, where there are just over two TA citations for every OA citation: a reversal of the above favouring TA articles.

Table 7.15 Citing to cited articles by region - % by row

				Cited Articles by Region												
OA				USA	Rest of Europe	UK	Rest of World	China	Pacific Rim	Canada	France	Germany	Italy	Japan	Spain	Total
Toll Access	Citing Articles by Region	USA	Count	115	18	16	7	9	7	13	16	14	9	4	3	231
			% within Region 2	49.8%	7.8%	6.9%	3.0%	3.9%	3.0%	5.6%	6.9%	6.1%	3.9%	1.7%	1.3%	100.0%
		Rest of Europe	Count	37	16	13	9	6	1	9	7	8	8	5	1	120
			% within Region 2	30.8%	13.3%	10.8%	7.5%	5.0%	.8%	7.5%	5.8%	6.7%	6.7%	4.2%	.8%	100.0%
		UK	Count	18	9	12	2	1	2	1	4	3	3	3	2	60
			% within Region 2	30.0%	15.0%	20.0%	3.3%	1.7%	3.3%	1.7%	6.7%	5.0%	5.0%	5.0%	3.3%	100.0%
		Rest of World	Count	17	9	2	5	3	2	1	0	2	5	7	2	55
			% within Region 2	30.9%	16.4%	3.6%	9.1%	5.5%	3.6%	1.8%	.0%	3.6%	9.1%	12.7%	3.6%	100.0%
		China	Count	32	4	3	9	20	5	1	4	0	3	2	7	90
			% within Region 2	35.6%	4.4%	3.3%	10.0%	22.2%	5.6%	1.1%	4.4%	.0%	3.3%	2.2%	7.8%	100.0%
		Pacific Rim	Count	15	4	2	6	2	7	2	4	4	4	1	0	51
			% within Region 2	29.4%	7.8%	3.9%	11.8%	3.9%	13.7%	3.9%	7.8%	7.8%	7.8%	2.0%	.0%	100.0%
		Canada	Count	13	1	0	1	5	1	1	1	0	2	2	0	27
			% within Region 2	48.1%	3.7%	.0%	3.7%	18.5%	3.7%	3.7%	3.7%	.0%	7.4%	7.4%	.0%	100.0%
		France	Count	24	5	3	3	1	0	0	14	3	4	0	2	59
			% within Region 2	40.7%	8.5%	5.1%	5.1%	1.7%	.0%	.0%	23.7%	5.1%	6.8%	.0%	3.4%	100.0%
		Germany	Count	24	9	4	1	4	0	3	4	11	4	5	1	70
			% within Region 2	34.3%	12.9%	5.7%	1.4%	5.7%	.0%	4.3%	5.7%	15.7%	5.7%	7.1%	1.4%	100.0%
		Italy	Count	23	1	7	4	1	5	2	12	6	16	0	4	81
			% within Region 2	28.4%	1.2%	8.6%	4.9%	1.2%	6.2%	2.5%	14.8%	7.4%	19.8%	.0%	4.9%	100.0%
		Japan	Count	16	1	4	1	1	4	0	0	0	1	24	0	52
			% within Region 2	30.8%	1.9%	7.7%	1.9%	1.9%	7.7%	.0%	.0%	.0%	1.9%	46.2%	.0%	100.0%
		Spain	Count	10	3	5	6	1	1	0	2	4	4	1	2	39
			% within Region 2	25.6%	7.7%	12.8%	15.4%	2.6%	2.6%	.0%	5.1%	10.3%	10.3%	2.6%	5.1%	100.0%
	Total		Count	344	80	71	54	54	35	33	68	55	63	54	24	935
			% within Region 2	36.8%	8.6%	7.6%	5.8%	5.8%	3.7%	3.5%	7.3%	5.9%	6.7%	5.8%	2.6%	100.0%
Open Access	Citing Articles by Region	USA	Count	359	62	34	11	14	20	20	37	18	20	3	12	610
			% within Region 2	58.9%	10.2%	5.6%	1.8%	2.3%	3.3%	3.3%	6.1%	3.0%	3.3%	.5%	2.0%	100.0%
		Rest of Europe	Count	133	55	27	9	9	5	7	32	19	19	2	4	321
			% within Region 2	41.4%	17.1%	8.4%	2.8%	2.8%	1.6%	2.2%	10.0%	5.9%	5.9%	.6%	1.2%	100.0%
		UK	Count	46	13	17	3	3	6	0	12	8	3	1	2	114
			% within Region 2	40.4%	11.4%	14.9%	2.6%	2.6%	5.3%	.0%	10.5%	7.0%	2.6%	.9%	1.8%	100.0%
		Rest of World	Count	31	22	4	5	5	2	3	12	9	7	1	3	104
			% within Region 2	29.8%	21.2%	3.8%	4.8%	4.8%	1.9%	2.9%	11.5%	8.7%	6.7%	1.0%	2.9%	100.0%
		China	Count	56	16	10	8	7	3	3	8	21	1	3	4	140
			% within Region 2	40.0%	11.4%	7.1%	5.7%	5.0%	2.1%	2.1%	5.7%	15.0%	.7%	2.1%	2.9%	100.0%
		Pacific Rim	Count	45	19	5	2	8	10	3	11	8	5	4	4	124
			% within Region 2	36.3%	15.3%	4.0%	1.6%	6.5%	8.1%	2.4%	8.9%	6.5%	4.0%	3.2%	3.2%	100.0%
		Canada	Count	31	8	4	0	0	1	5	3	5	2	0	1	60
			% within Region 2	51.7%	13.3%	6.7%	.0%	.0%	1.7%	8.3%	5.0%	8.3%	3.3%	.0%	1.7%	100.0%
		France	Count	53	14	9	9	1	0	3	29	8	8	0	3	137
			% within Region 2	38.7%	10.2%	6.6%	6.6%	.7%	.0%	2.2%	21.2%	5.8%	5.8%	.0%	2.2%	100.0%
		Germany	Count	73	21	12	6	1	11	4	28	46	8	2	2	214
			% within Region 2	34.1%	9.8%	5.6%	2.8%	.5%	5.1%	1.9%	13.1%	21.5%	3.7%	.9%	.9%	100.0%
		Italy	Count	35	9	6	6	1	2	0	19	11	27	0	4	120
			% within Region 2	29.2%	7.5%	5.0%	5.0%	.8%	1.7%	.0%	15.8%	9.2%	22.5%	.0%	3.3%	100.0%
		Japan	Count	46	16	3	1	1	0	3	4	6	1	6	3	90
			% within Region 2	51.1%	17.8%	3.3%	1.1%	1.1%	.0%	3.3%	4.4%	6.7%	1.1%	6.7%	3.3%	100.0%
		Spain	Count	20	6	3	3	0	2	1	9	10	3	3	3	63
			% within Region 2	31.7%	9.5%	4.8%	4.8%	.0%	3.2%	1.6%	14.3%	15.9%	4.8%	4.8%	4.8%	100.0%
	Total		Count	928	261	134	63	50	62	52	204	169	104	25	45	2097
			% within Region 2	44.3%	12.4%	6.4%	3.0%	2.4%	3.0%	2.5%	9.7%	8.1%	5.0%	1.2%	2.1%	100.0%

Table 7.16 Citing to cited articles by region - % by column

				Cited Articles by Region												
OA				USA	Rest of Europe	UK	Rest of World	China	Pacific Rim	Canada	France	Germany	Italy	Japan	Spain	Total
Toll Access	Citing Articles by Region	USA	Count	115	18	16	7	9	7	13	16	14	9	4	3	231
			% within Region	33.4%	22.5%	22.5%	13.0%	16.7%	20.0%	39.4%	23.5%	25.5%	14.3%	7.4%	12.5%	24.7%
		Rest of Europe	Count	37	16	13	9	6	1	9	7	8	8	5	1	120
			% within Region	10.8%	20.0%	18.3%	16.7%	11.1%	2.9%	27.3%	10.3%	14.5%	12.7%	9.3%	4.2%	12.8%
		UK	Count	18	9	12	2	1	2	1	4	3	3	3	2	60
			% within Region	5.2%	11.3%	16.9%	3.7%	1.9%	5.7%	3.0%	5.9%	5.5%	4.8%	5.6%	8.3%	6.4%
		Rest of World	Count	17	9	2	5	3	2	1	0	2	5	7	2	55
			% within Region	4.9%	11.3%	2.8%	9.3%	5.6%	5.7%	3.0%	.0%	3.6%	7.9%	13.0%	8.3%	5.9%
		China	Count	32	4	3	9	20	5	1	4	0	3	2	7	90
			% within Region	9.3%	5.0%	4.2%	16.7%	37.0%	14.3%	3.0%	5.9%	.0%	4.8%	3.7%	29.2%	9.6%
		Pacific Rim	Count	15	4	2	6	2	7	2	4	4	4	1	0	51
			% within Region	4.4%	5.0%	2.8%	11.1%	3.7%	20.0%	6.1%	5.9%	7.3%	6.3%	1.9%	.0%	5.5%
		Canada	Count	13	1	0	1	5	1	1	1	0	2	2	0	27
			% within Region	3.8%	1.3%	.0%	1.9%	9.3%	2.9%	3.0%	1.5%	.0%	3.2%	3.7%	.0%	2.9%
		France	Count	24	5	3	3	1	0	0	14	3	4	0	2	59
			% within Region	7.0%	6.3%	4.2%	5.6%	1.9%	.0%	.0%	20.6%	5.5%	6.3%	.0%	8.3%	6.3%
		Germany	Count	24	9	4	1	4	0	3	4	11	4	5	1	70
			% within Region	7.0%	11.3%	5.6%	1.9%	7.4%	.0%	9.1%	5.9%	20.0%	6.3%	9.3%	4.2%	7.5%
		Italy	Count	23	1	7	4	1	5	2	12	6	16	0	4	81
			% within Region	6.7%	1.3%	9.9%	7.4%	1.9%	14.3%	6.1%	17.6%	10.9%	25.4%	.0%	16.7%	8.7%
		Japan	Count	16	1	4	1	1	4	0	0	0	1	24	0	52
			% within Region	4.7%	1.3%	5.6%	1.9%	1.9%	11.4%	.0%	.0%	.0%	1.6%	44.4%	.0%	5.6%
		Spain	Count	10	3	5	6	1	1	0	2	4	4	1	2	39
			% within Region	2.9%	3.8%	7.0%	11.1%	1.9%	2.9%	.0%	2.9%	7.3%	6.3%	1.9%	8.3%	4.2%
	Total		Count	344	80	71	54	54	35	33	68	55	63	54	24	935
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Open Access	Citing Articles by Region	USA	Count	359	62	34	11	14	20	20	37	18	20	3	12	610
			% within Region	38.7%	23.8%	25.4%	17.5%	28.0%	32.3%	38.5%	18.1%	10.7%	19.2%	12.0%	26.7%	29.1%
		Rest of Europe	Count	133	55	27	9	9	5	7	32	19	19	2	4	321
			% within Region	14.3%	21.1%	20.1%	14.3%	18.0%	8.1%	13.5%	15.7%	11.2%	18.3%	8.0%	8.9%	15.3%
		UK	Count	46	13	17	3	3	6	0	12	8	3	1	2	114
			% within Region	5.0%	5.0%	12.7%	4.8%	6.0%	9.7%	.0%	5.9%	4.7%	2.9%	4.0%	4.4%	5.4%
		Rest of World	Count	31	22	4	5	5	2	3	12	9	7	1	3	104
			% within Region	3.3%	8.4%	3.0%	7.9%	10.0%	3.2%	5.8%	5.9%	5.3%	6.7%	4.0%	6.7%	5.0%
		China	Count	56	16	10	8	7	3	3	8	21	1	3	4	140
			% within Region	6.0%	6.1%	7.5%	12.7%	14.0%	4.8%	5.8%	3.9%	12.4%	1.0%	12.0%	8.9%	6.7%
		Pacific Rim	Count	45	19	5	2	8	10	3	11	8	5	4	4	124
			% within Region	4.8%	7.3%	3.7%	3.2%	16.0%	16.1%	5.8%	5.4%	4.7%	4.8%	16.0%	8.9%	5.9%
		Canada	Count	31	8	4	0	0	1	5	3	5	2	0	1	60
			% within Region	3.3%	3.1%	3.0%	.0%	.0%	1.6%	9.6%	1.5%	3.0%	1.9%	.0%	2.2%	2.9%
		France	Count	53	14	9	9	1	0	3	29	8	8	0	3	137
			% within Region	5.7%	5.4%	6.7%	14.3%	2.0%	.0%	5.8%	14.2%	4.7%	7.7%	.0%	6.7%	6.5%
		Germany	Count	73	21	12	6	1	11	4	28	46	8	2	2	214
			% within Region	7.9%	8.0%	9.0%	9.5%	2.0%	17.7%	7.7%	13.7%	27.2%	7.7%	8.0%	4.4%	10.2%
		Italy	Count	35	9	6	6	1	2	0	19	11	27	0	4	120
			% within Region	3.8%	3.4%	4.5%	9.5%	2.0%	3.2%	.0%	9.3%	6.5%	26.0%	.0%	8.9%	5.7%
		Japan	Count	46	16	3	1	1	0	3	4	6	1	6	3	90
			% within Region	5.0%	6.1%	2.2%	1.6%	2.0%	.0%	5.8%	2.0%	3.6%	1.0%	24.0%	6.7%	4.3%
		Spain	Count	20	6	3	3	0	2	1	9	10	3	3	3	63
			% within Region	2.2%	2.3%	2.2%	4.8%	.0%	3.2%	1.9%	4.4%	5.9%	2.9%	12.0%	6.7%	3.0%
	Total		Count	928	261	134	63	50	62	52	204	169	104	25	45	2097
			% within Region	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

In every case, the majority of the citations go to the USA for TA and OA citations from other regions. Four countries take the highest number of citations for both OA and TA articles. For OA articles, this is the USA first, followed by Europe, Germany and China, whilst for TA articles Germany and China's positions are reversed. Table 7.17 shows the citing country and the income group of the related cited articles.

Table 7.17 Citing country to cited income group

OA				Cited Income				Total
				Low	Lower middle	Upper middle	High	
Toll Access	Citing Country	USA	Count	0	17	3	211	231
			% within Region	.0%	7.4%	1.3%	91.3%	100.0%
		Rest of Europe	Count	0	7	10	103	120
			% within Region	.0%	5.8%	8.3%	85.8%	100.0%
		UK	Count	0	2	0	58	60
			% within Region	.0%	3.3%	.0%	96.7%	100.0%
		Rest of World	Count	0	3	9	43	55
			% within Region	.0%	5.5%	16.4%	78.2%	100.0%
		China	Count	3	21	6	60	90
			% within Region	3.3%	23.3%	6.7%	66.7%	100.0%
		Pacific Rim	Count	0	5	2	44	51
			% within Region	.0%	9.8%	3.9%	86.3%	100.0%
		Canada	Count	0	5	1	21	27
			% within Region	.0%	18.5%	3.7%	77.8%	100.0%
		France	Count	0	2	2	55	59
			% within Region	.0%	3.4%	3.4%	93.2%	100.0%
		Germany	Count	0	5	1	64	70
			% within Region	.0%	7.1%	1.4%	91.4%	100.0%
		Italy	Count	2	6	3	70	81
			% within Region	2.5%	7.4%	3.7%	86.4%	100.0%
		Japan	Count	0	3	1	48	52
			% within Region	.0%	5.8%	1.9%	92.3%	100.0%
		Spain	Count	0	3	4	32	39
			% within Region	.0%	7.7%	10.3%	82.1%	100.0%
	Total		Count	5	79	42	809	935
			% within Region	.5%	8.4%	4.5%	86.5%	100.0%
Open Access	Citing Country	USA	Count	0	14	12	584	610
			% within Region	.0%	2.3%	2.0%	95.7%	100.0%
		Rest of Europe	Count	0	11	7	303	321
			% within Region	.0%	3.4%	2.2%	94.4%	100.0%
		UK	Count	0	3	2	109	114
			% within Region	.0%	2.6%	1.8%	95.6%	100.0%
		Rest of World	Count	1	5	3	95	104
			% within Region	1.0%	4.8%	2.9%	91.3%	100.0%
		China	Count	0	7	4	129	140
			% within Region	.0%	5.0%	2.9%	92.1%	100.0%
		Pacific Rim	Count	0	12	0	112	124
			% within Region	.0%	9.7%	.0%	90.3%	100.0%
		Canada	Count	0	0	0	60	60
			% within Region	.0%	.0%	.0%	100.0%	100.0%
		France	Count	0	1	1	135	137
			% within Region	.0%	.7%	.7%	98.5%	100.0%
		Germany	Count	0	3	5	206	214
			% within Region	.0%	1.4%	2.3%	96.3%	100.0%
		Italy	Count	0	1	3	116	120
			% within Region	.0%	.8%	2.5%	96.7%	100.0%
		Japan	Count	1	1	1	87	90
			% within Region	1.1%	1.1%	1.1%	96.7%	100.0%
		Spain	Count	0	0	0	63	63
			% within Region	.0%	.0%	.0%	100.0%	100.0%
	Total		Count	2	58	38	1999	2097
			% within Region	.1%	2.8%	1.8%	95.3%	100.0%

Figure 7.60 illustrates the data from Table 7.17 by percentage of the whole count for each category the OA status of the citations to each region. In the case of the USA, it has 24.7% of TA citations (231/935) and 29.1% of OA citations (610/2097). Of the twelve regions, only five have a greater overall percentage of OA than TA citations even though in every case each region had a greater number of OA than TA citations. It is noticeable that China and the Rest of World have a greater percentage of their articles that are toll access.

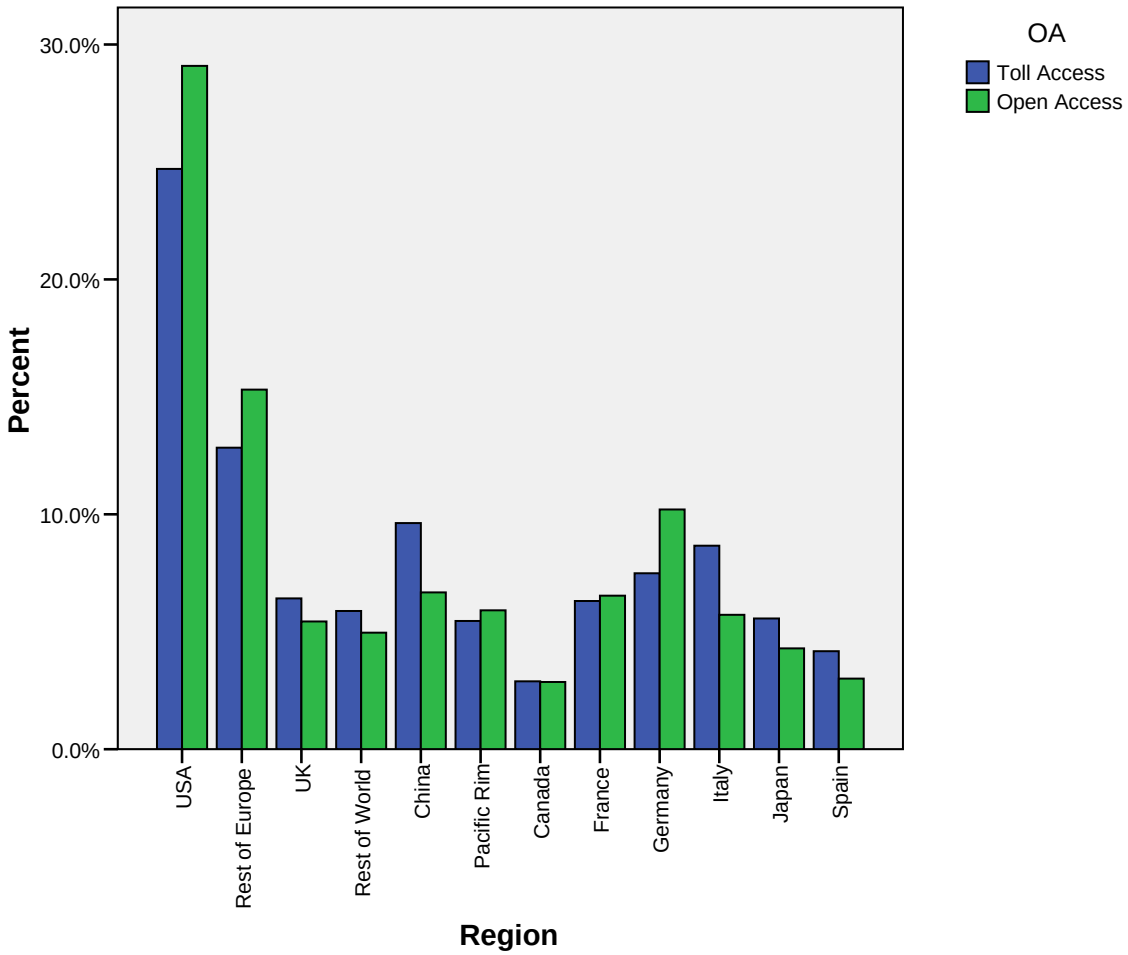


Figure 7.60 Percentage of citations by OA status

7.33. Logistic regression analysis – collective subjects

The records collected in the first round of data collection were categorised into those articles that were OA or TA. For the purposes of logistic regression, this OA/TA outcome is classed as a dependent categorical dichotomy. The associated bibliographic data collected with these records are potential (independent) predictor variables of the OA status of these articles. Logistic regression can use these predictor variables to attempt to predict the categorical

outcome of an article's OA status. SPSS software was used to run a forward stepwise likelihood ratio logistic regression analysis of the 4633 records found in the first round of data collection to determine if the observed frequency of OA articles can be better predicted by using these predictor variables. In this model the OA status of an article was coded 0 = TA (Toll Access) and 1 = OA (Open Access).

To check for multicollinearity, the linear correlation of the independent variables were tested using the Pearson r correlation coefficient. Several of the variables had a relatively high correlation, but these were variables that were later discarded from the model during processing. Appendix G lists the 18 independent variables that were entered into the model. After calculation ten of these variables had a significance level greater than 0.05 and hence were discarded by the model as having less predictive power than the remaining eight. By iterative addition through eight steps, the remaining significant variables were added to the regression model as having a predictive effect on determining the OA status of an article:

- Subject - sociology
- Country of origin by first author affiliation - USA
- Total citations to an article
- Article subject - ecology
- Impact factor of the journal in which the article appeared
- Country of origin by first author affiliation - Europe
- Number of authors per article
- Other author citations appearing in the same journal as the original article

Some of the tabular output from the regression model produced by SPSS are shown below in Table 7.18 to Table 7.24. Table 7.18 gives the number of article records from the original data; the model simply predicts initial class membership based on the most frequently occurring category; in this case TA articles. The initial -2 log likelihood value for the baseline model was 6418.841. As shown in Table 7.19 after adding the eight predictive independent variables to the model this drops to 5584.515. A change of 834.326, this indicates the model is predicting category outcome more accurately. The statistical significance of this change is indicated by the Chi-squared result ($\chi^2(8) = 834.326$, $p<0.001$)

this is shown in the *Omnibus Tests of Model Coefficients* within Table 7.20 which confirms that the overall model is significant. The *Hosmer and Lemeshow's* goodness of fit test entry in Table 7.21 gives a Chi-squared result of ($\chi^2(8) = 8.680$, $p < 0.370$). In contrast to the earlier result, this non-significant outcome ($p > 0.05$) indicates a good fit with the observed data (Pallant 2005, p.167).

When the second classification table, (Table 7.22) is compared to the first (Table 7.18); it shows the model after the eight variables have been added and indicates a modest improvement to 67.2% in the success of the model at predicting category membership, an overall improvement of 16.4%.

The results shown in Table 7.23 shows the contribution of each of the variables to the total -2 log likelihood value and their significance to the model. Of these, the variables sociology and ecology contribute approximately 75% of the model's improvement. The Wald statistic shown in Table 7.24 illustrates the contribution of each predictor in the model. With their statistical significance below $p < 0.05$, all of the variables are significant. The 'B' values given in the table are both positive and negative and indicate the direction of the relationship (which factors increase the probability of an OA or TA status). In the case of sociology, this appears as a negative value and a "…negative B value indicate[s] that an increase in the independent variable score will result in a decreased probability of the case recording a score of one in the dependent variable" [being OA in this case] (Pallant 2005, p.169). The Exp(B) value indicates the odds ratio for each of the variables. Tabachnick and Fidel (2001, p.548) define the "…odds ratio [as] the increase (or decrease if the ratio is less than one) in the odds of being in one outcome category when the value of the predictor increases by one unit". The values of Exp(B) given in Table 7.24 confirm that sociology, ecology and 'other author' (other author citations from articles appearing in the same journal) are less than one and have odds that will decrease the likelihood of an OA outcome as their predictor value increases. The 'other author' variable is very small unlike the variable 'Citations from other authors in other journals', which has the majority of the other author citations. Conversely increases in the predictor variables, number of authors, impact factor and total citations will result in an increase in odds of the outcome (OA) occurring. These odds are however only just over one and hence have only a very small effect (Field 2005, p.241).

Pallant (2005, p.168) suggests that the positive and negative predictive values may be calculated. The predictive values are the percentage of cases that the model actually correctly classifies as having the characteristic for that group. From the second classification table the predicted and observed value of 1674 OA articles is divided by the total in the open access column (1674/913+1674) to give a positive predictive value of 64.71%. For the negative predictive value the same calculation is undertaken but for the observed value of 1438 in the toll access column (1438/606+1438) to give 70.35%. The model it appears is more successful at identifying TA articles than OA articles.

Table 7.18 Classification table for regression model OA/NOA

Classification Table[a,b]

Observed			Predicted		
			OA		Percentage Correct
			Toll Access	Open Access	
Step 0	OA	Toll Access	2351	0	100.0
		Open Access	2280	0	.0
	Overall Percentage				50.8

a. Constant is included in the model.

b. The cut value is .500

Table 7.19 Iteration history

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients								
			Constant	Sociology	USA	Totcites	Ecology	IPFactor	Europe	Numberof authors	OA
Step 8	1	5614.396	-.367	-1.627	.627	.038	-.934	.074	.273	.058	-.040
	2	5584.880	-.571	-1.874	.752	.056	-1.146	.099	.342	.073	-.062
	3	5584.515	-.595	-1.898	.762	.058	-1.174	.103	.348	.074	-.065
	4	5584.515	-.596	-1.898	.762	.058	-1.175	.103	.348	.074	-.065

a. Method: Forward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 6418.841

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Table 7.20 Omnibus results

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 8	Step	5.021	1	.025
	Block	834.326	8	.000
	Model	834.326	8	.000

Table 7.21 Hosmer and Lemeshow results

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
8	8.680	8	.370

Table 7.22 Final classification table at step eight

Classification Table[a]

			Predicted		
			OA		Percentage Correct
	Observed		Toll Access	Open Access	
Step 8	OA	Toll Access	1438	913	61.2
		Open Access	606	1674	73.4
	Overall Percentage				67.2

a. The cut value is .500

Table 7.23 Model if term removed

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 8	Numberofauthors	-2796.864	9.214	1	.002
	IPFactor	-2799.967	15.420	1	.000
	Totcites	-2851.756	118.997	1	.000
	OA	-2794.768	5.021	1	.025
	USA	-2836.162	87.809	1	.000
	Europe	-2798.748	12.981	1	.000
	Ecology	-2842.786	101.057	1	.000
	Sociology	-3054.050	523.586	1	.000

Table 7.24 Variables left in equation after eight steps

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 8[h]	Numberofauthors	.074	.025	9.048	1	.003	1.077	1.026	1.129
	IPFactor	.103	.027	14.214	1	.000	1.109	1.051	1.170
	Totcites	.058	.006	101.551	1	.000	1.060	1.048	1.072
	OA	-.065	.029	5.041	1	.025	.937	.885	.992
	USA	.762	.082	85.545	1	.000	2.143	1.823	2.519
	Europe	.348	.097	12.929	1	.000	1.417	1.172	1.713
	Ecology	-1.175	.120	95.277	1	.000	.309	.244	.391
	Sociology	-1.898	.089	450.333	1	.000	.150	.126	.179
	Constant	-.596	.109	30.139	1	.000	.551		

a. Variable(s) entered on step 1: Sociology.

b. Variable(s) entered on step 2: USA.

c. Variable(s) entered on step 3: Totcites.

d. Variable(s) entered on step 4: Ecology.

e. Variable(s) entered on step 5: IPFactor.

f. Variable(s) entered on step 6: Europe.

g. Variable(s) entered on step 7: Numberofauthors.

h. Variable(s) entered on step 8: OA.

Appendix H gives the output for the same first round data but as input to SPSS in the 'enter' and 'backward stepwise' mode. The results for the 'enter' model show a minor difference in the -2 log likelihood result after four iterations to 5577.461 a difference of seven over the forward stepwise entry model. All results are significant. There is a minor difference in the model's correct prediction of category membership of 0.2% percentage points compared to the forward stepwise entry model. What is noticeable, however, is the exclusion of sociology as a predictive variable and the inclusion of economics and applied maths as the major contributors to the model's predictive power with very strong odds of increasing the likelihood of an articles' OA status. Europe and the USA are included in the module as adding to the models predictive power. What is evident is the dominance of the subject in predicting the categorical membership of an article by its subject both in this model and in the forward stepwise results. Similarly examination of the 'backward stepwise' SPSS output in Appendix H shows an almost identical result to the 'enter' model.

7.34. Logistic regression analysis – individual subjects

The results shown in Table 7.25 are the primary results from a logistic model for individual subjects from the first round of data collection.

Table 7.25 Logistic regression results at subject level – first round data

Subject	Initial -2 log likelihood	Model -2 log likelihood	Baseline model prediction	Model prediction	Statistically significant. good fit	Dominant predictor variable	Direction
Sociology	1196.553	1124.611	78.9	80.1	Yes	Other author citations	Positive
Maths	1571.307	1497.509	58.5	61.2	Yes	Total citations	Positive
Economics	1480.727	1411.765	64.8	65.3	Yes	Other author citations	Positive
Ecology	1619.741	1502.344	52.8	62.6	Yes	USA	Positive

Forward stepwise entry method

Apart from a modest gain in category prediction for ecology, the results show very small differences in the -2 log likelihood statistic and category prediction. All results are

significant and positive, with only ecology having a dominant predictor variable not related to citation counts. In all cases, the next dominant predictor variable is country of origin by first author affiliation.

7.35. Logistic regression analysis – second round individual subjects

Examination at subject level using second round data is shown in Table 7.26. First and second round data for sociology are comparable with the second round data taken a year later. The primary records in the maths sample are the same, but have been inflated by the addition of citer details and so cannot be compared to the first round data. Similarly, records for economics were taken from lower impact journals and the sample for ecology was entirely random. The results indicate a relatively minor explanation of the variables as indicated in the small change in the -2 log-likelihood statistic from the initial baseline model to the model after predictor variables have been added.

Table 7.26 Logistic regression results at subject level – second round data

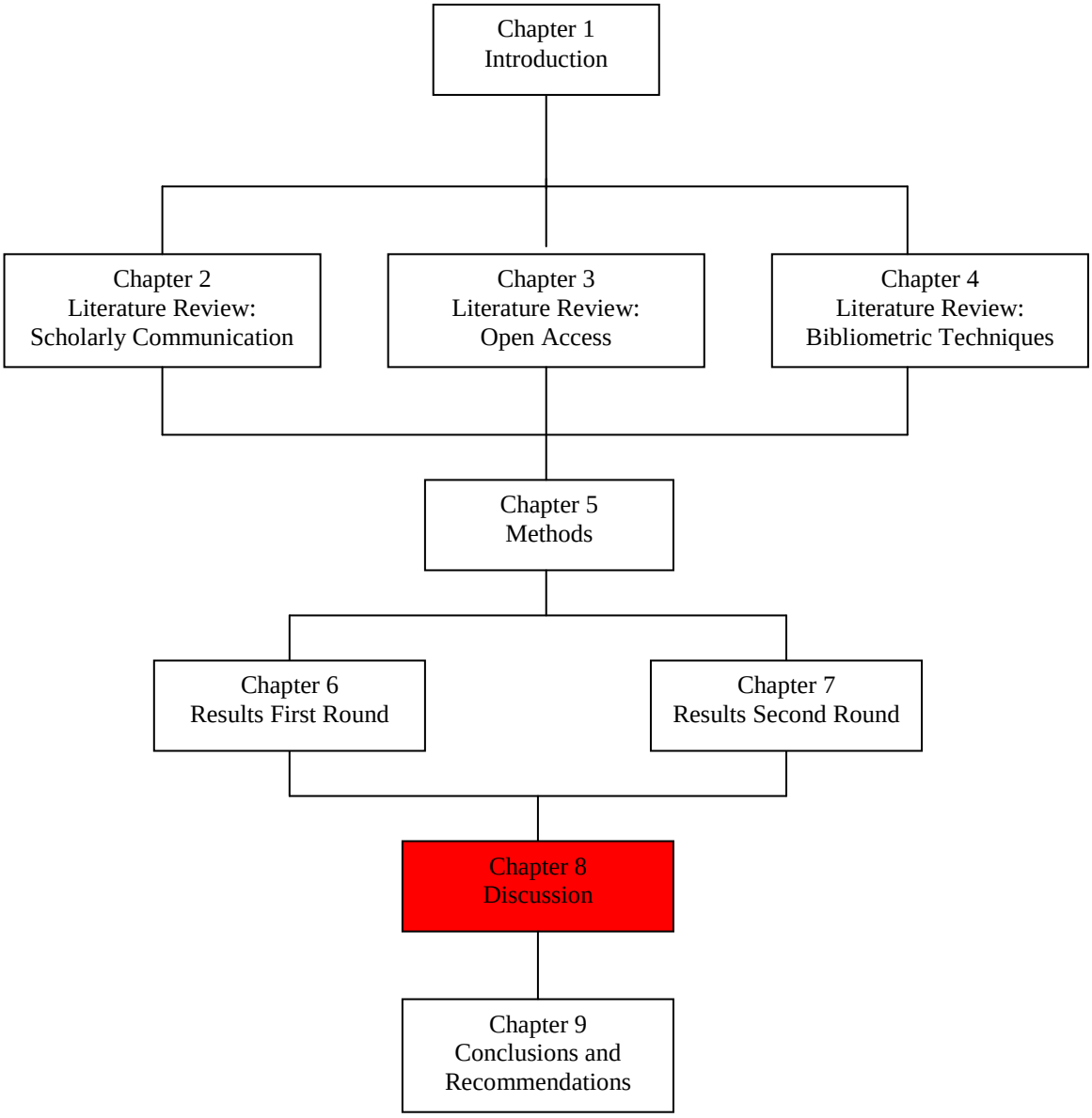
Subject	Initial -2 log likelihood	Model -2 log likelihood	Baseline model prediction	Model prediction	Statistically significant. good fit	Dominant predictor variable	Direction
Sociology	1026.361	970.040	75.7	76.3	Yes	Total citations	Positive
Maths	4293.377	3910.791	68.2	68	No	Other author citations	Positive
Economics	1333.065	1273.528	54.4	60.1	No	Total citations	Positive
Ecology	805.409	681.763	65.7	72.7	Yes	Impact factor	Positive

Forward stepwise entry method

Differences between the baseline model of category membership and after predictor variables have been added, are also relatively small. In the case of maths and economics, the *Hosmer and Lemeshow Test* reports that these two subjects are a poor fit with the baseline data. In the case of maths, the model included the country of origin of those citing an article and their country's income status. These two predictor variables were discarded by the logistic regression model as non-explanatory variables. The case for those citing OA articles

based on only doing so because the articles are freely available in countries that are ‘poor’ and so increasing the odds of an article being OA appears weak. Only in ecology as in the first round data is there a modest improvement in the model’s predictive power but with a change of the dominant predictor to impact factor.

Chapter 8 Discussion



8.1. Introduction

This chapter is broadly split into the following six basic themes:-

- Introduction
- Open access – overall results
- Citation counts, correlation and associations
- Authorship levels, subjects and country of origin
- OA article discovery and
- Causation.

Each theme is briefly outlined and then discussed where appropriate, within the context of the literature review, and the aims and objectives. The primary research question on citation advantage and its first four related objectives, including the secondary question of causation are covered. The chapter concludes with a summary, linking together the different themes together with some thoughts on further research and what the results might mean for scholarly communication.

8.1.1. Background, aims and objectives

OA articles are said to accrue more citations than comparable TA articles. The aim of the research was to assess whether OA articles do, in fact, receive more citations than comparable TA articles and in addition, if this was so, to determine whether OA status was solely or partly the cause of this advantage.

Article records from four disciplines were collected and their citation counts analysed for evidence of any OA/TA citation advantage, and to identify if any characteristics of these two sets of articles might explain, if at all, the causal links between them and any citation advantage. The OA status of articles was determined by searching for freely accessible versions of them on the World Wide Web using different search tools. Two rounds of data collection were undertaken using the same subjects but with different sampling regimes to meet the stated objectives of the work. The results were subject to

statistical analysis. The level of open access and the citation advantage varied between subjects and differences were evident between the two rounds of data. Overall, the findings suggest that a citation advantage was consistently evident for those articles that were OA, but the causes of that advantage were unclear.

8.1.2. Research questions

The following research questions were posed:

- Do OA articles receive more citations than comparable TA articles and thus have greater research impact?
- Is there an identifiable causal link between any citation advantage and an article's OA status?

This gave rise to the following null hypotheses:

- That there is no significant difference in the rates of citation between open access and toll access articles in favour of open access articles, in a group of subjects over a fixed period.
- That there are no causal links between the open access status of an article and the collective bibliographic details associated with that status.

To test the validity of these hypotheses, five objectives were set:

Objective 1: Determine the OA citation advantage or otherwise by examining the citation counts of high impact journal articles from four discrete disciplines.

Objective 2: Confirm that the results found in Objective 1 were not a chance event.

Objective 3: Ascertain whether the OA/TA citation advantage is randomly evident in a population of journal articles and whether there is an early access advantage evident from OA articles in terms of patterns of earlier citations.

Objective 4: Determine if lower impact journals have a similar distribution of OA/TA articles and citation characteristics to their high impact counterparts.

Objective 5: To determine if causal links can be found between the OA/TA status of an article and the collective bibliographic details of the records associated with that status.

8.1.3. Areas of interest

There are some results in the research that are of particular interest and these will be discussed within the above themes. Amongst these was the examination of those authors, who cited the applied maths articles, by their country of origin and the economic status of that country to see if any relationship could be established between those accessing OA articles and their perceived general ability to access high impact journals. In addition, the use of logistic regression to try to identify the causes of any OA citation advantage and the differing success rates of the search tools used to find OA articles. Additionally there is scope to compare the results between the two rounds of data collection where it is evident that the two samples exhibit different characteristics and levels of OA.

8.2. Open access - overall results

Lawrence's seminal study, started from the premise that "usage increases when access is more convenient" (2001, p.521). Lawrence found that conference articles that are freely available generally have higher citation counts than those that are not, but could not say, with any certainty what the cause of the correlation was between citation rates and online availability. Similarly, Kurtz (2004) found that access policies to journals could reduce the frequency with which full text articles were accessed by up to a factor of two. Increasing access, he suggested, increases both the likelihood that an article will be read and that it will be cited. The results found here are consistent with the initial view that the more accessible an article is, the more likely it will be cited and the more likely it will have a greater mean citation count than TA articles. The four subjects chosen here applied maths, sociology, ecology and economics show a consistent citation advantage when the mean citations counts of OA articles are compared to those of TA articles. As shown in Table 6.1 and Table 6.2 the gross advantage varied from 44% for ecology to 88% for sociology and increased, when all self-citations were excluded, to 49% and 103% respectively. This advantage is also evident from the data collected for objectives 2-4 for the second round of data collection. In this second round data,

sociology had a gross advantage of 80% rising to 94% when self-citations were excluded. Similarly, the results from ecology were respectively 105% and 135% and for economics, the figures were 74% and 82%. Taken collectively, all of the results consistently report an advantage for OA articles irrespective of subject, even, in the case of the second round economics results, where articles were taken from lower impact journals.

8.2.1. Levels of open access

When the actual levels of OA are examined for each of the subjects, there is considerable variation, with economics having a greater level of OA than the other subjects. In the case of applied maths the level of open access was 59%, this is in contrast to Antelman (2004, p.377) who using *Google* alone found from 608 mathematics articles an open access rate of 69%. She noted in her preliminary sample of 50 articles that 60% of the mathematics articles had been found at a subject repository, although not recorded it is highly likely that this would have been arXiv or Citebase where maths authors are known depositors. The likely explanation for this marked difference of ten percentage points is the difference between the pure and applied maths where self-archiving is, at arXiv for example, dominated by pure maths subjects rather than applied mathematical subjects and is part of the preprint culture associated with this archive. Antelman (2004) also noted that the subject, with its relatively high level of open access, had the highest citation advantage at 91%. Although not directly comparable, applied maths here was only third in its citation advantage but second in its level of open access. Whilst arXiv is a notable repository for mathematics, Davis and Fromerth (2007), who examined 2765 articles from four mathematics journals could only find 18.5% of them in the repository, in marked contrast to the level of open access given above. It is assumed that this discrepancy lies in part to the fact that the authors did not search for OA versions of the articles beyond arXiv.

Much of the research into any OA advantage is based on comparing the article records found in the arXiv repository. Its content is mainly in maths and physics. However, work by Hajjem, Harnad and Gingras (n.d.) covered ten disciplines, including sociology and economics. They found levels of open access respectively of 16% and 13.5%. Whilst in this research, sociology had the smallest proportion of OA articles, there was a

small increase in the level of open access to sociology in the second round data. The increase was from 21% to 24.4%. This could be just chance or perhaps could be attributed to the growing awareness of OA and the growth in the number of institutional repositories, especially when it is noted that the percentage share of the OA articles increased for the UK and Europe by about three percentage points in the second round data at the expense of North America, whilst the RoW remained static. For this subject, this is a useful comparable result, given that the Hajjem, Harnad and Gingras sample that made up the 16% was drawn from a much larger range of journals and impact factors. This result is interesting given that the second series of article records collected here were taken one year later, although there was some variation in the journals from which the articles were taken. Hajjem, Harnad and Gingras (n.d.) also found a slow but steady increase in the number of articles that were OA and that this increase was more apparent in the more highly cited articles.

Differing levels of open access are a feature of the research. The rate of open access was the highest for economics at 65% and this ties in with the work reported by Bergstrom and Lavaty (2007), who examined how often economists, of different standing self-archived from a mixed range of 33 economics journals. From a sample of two issues from each of these journals, they found that the overall rate of self-archiving was about 70%. This is not that dissimilar to the 65% and 54% found here, respectively, from the first round high impact journals and second round data drawn from mid-range impact journals. When split by the more highly and less cited journals, the self-archiving ranged from 90% for the most cited and 50% for the less cited journals. The authors also examined the self-archiving rates within two issues of political science journals, a subject which Moed (2005, p.130) groups with the social sciences and which includes sociology as having similar communication characteristics and ISI coverage. Bergstrom and Lavaty (2007) found a self-archiving rate of 30% for the political science journals, which is fairly close to the first and second round results for sociology of 21% and 24% respectively. This result is consistent with Antelman's (2004) sample for political science, which was 29%.

The 13.5% rate of open access for economics articles found by Hajjem, Harnad and Gingras (n.d.) is different to the results found here, which was the highest at 65% or

with the 70% rate found by Bergstrom and Lavaty (2007). A possible explanation for this difference may be that the search for OA articles using a range of search tools did not trawl the distributed RePEc archives. However, the article sample taken by the three authors was from between 1992-2003 and RePEc in 1999 held about 70,000 records, compared to over half a million in 2008. If Bergstrom and Lavaty are correct in their assumption that authors writing in high impact economics journals now self-archive as a matter of routine, then it is conceivable that economics authors in the period in question simply self-archived less frequently, and/or were less aware of RePEc.

However, in an explanation of a developing trend towards higher self-archiving, Bergstrom and Lavaty (2007) concluded that self-archiving is the norm for research intensive authors from more prestigious institutions who publish in journals of higher impact, but less so for authors who publish in lower impact journals. They ascribed the higher rate of self-archiving by these research intensive authors to more influential journals, as an attempt to garner citations to enhance their income expectations. Perhaps more speculatively, the authors felt that it is the culture and norms of their peers in the institution in which authors work that is the most influential factor in whether an author self archives their work or not. This could possibly explain why the OA rate for the sociology sample Hajjem, Harnad and Gingras (n.d) found (16%) was relatively close to the 20% found here, in that authors in this discipline do not have a similar place to self-archive their work, are less driven by financial rewards and generally use a local repository or their own homepages.

Given that the sample was taken from mid-range journals, the second round sample for economics had a surprisingly high level of OA at 54%. This was a similar result to that of sociology, where the change between first and second round data was an increase in the level of OA in favour of authors from the UK and Europe. The drop from an OA level of 65% in the first round economics sample to 54% in the second is explained by the drop from 70% to 50% of first author affiliations from North America to the advantage of Europe and the RoW, whose share of the OA articles almost doubled with a small increase for the UK. When the overall level of first author affiliation is taken for North America for both OA and TA articles and compared for the economics articles, the percentage share drops from 65% for the first round data to 48% in the second. Not

only is there is a drop in the number of articles from North America for this mid-range sample, there is also a drop in the percentage of those articles that are OA from the first round data of 70% to 56% in the second. Interestingly, however, although the overall share of OA articles rises for the other regions, it is only in the UK that there is there a marginal increase in the percentage of OA articles when compared to the first round data. Clearly, in this second round data there is a steep drop in the number of articles from North America, a disproportionate rise in the number of TA articles from the region and a complementary rise in OA articles from the other territories with a limited change in the actual individual percentage share of OA articles. This reflects nicely the clear dominance of North American authors in high impact journals. It also shows that the lower the impact factor, at least in this sample, the fewer North American authors there are and, given their propensity to self-archive, the smaller the OA advantage.

The first round ecology data exhibits characteristics of a science subject with higher than average author counts and a greater level of citation than the other subjects. The two rounds of data collection for ecology were quite different with a purposive sample for the first and a random sample for the second; although different, there are evident points of comparison. Levels of OA were different, with the first round at 53% and the second at 34%. It is apparent, from the general results, that the greater the number of authors, the greater the likelihood that an article will be OA. On this basis it ought to be the case that ecology with 3.25 authors per paper for the first round data ought to have the highest level of OA. However, the subject is only ranked third ahead of sociology. Interestingly the author count for the second round data for the OA articles is even higher at 3.4 authors per paper. Unlike mathematics and economics, there are no mature subject repositories for ecology articles.

Davis and Connolly (2007) in a review of the use of the institutional repository at Cornell University found that academics had little knowledge and motivation to use it, preferring to use personal web pages or disciplinary repositories. The latter, they thought, had greater "community salience than one's affiliate institution". The authors also interviewed faculty staff, some of whom were from the Ecology and Evolution department. These staff thought that their peers had adequate access to the literature and those that did not could contact them directly or visit their personal web page. Although

these results are from one institution only, they point to a culture of non-disciplinary archiving and at best local self-archiving to a personal webpage. This, it is suggested, is the likely reason why OA rates in ecology are less than economics or applied mathematics, where there is more of a preprint culture and there are successful disciplinary repositories. Like the other subjects, ecology is dominated by first affiliation authors from North America. For the purposive sample, this was 53% of all articles and for the random sample, this was 39%. The distribution of first affiliation authors in the random sample moved away from the dominance of North America to the benefit of Europe and the RoW, with the UK also dropping its share from 12.4% to 8.4%. This ties in with the results from the other subjects, again showing how high impact journal articles are more likely to be OA and have a greater number of authors from North America. As discussed later in section 8.4.3, there is a clear indication that the higher the impact factor, the more likely the article will be OA in this subject with a clearly different distribution of TA articles that is more evenly spread amongst the different impact factors.

8.3. Citations, self-citations, correlations and associations

Citations are by definition the measure of any citation advantage. Their type and frequency varies with the type of discipline that the citing authors are working in, and to a certain extent, the level of self-citation varies depending on journal impact. Clearly, articles are either cited or un-cited and the level of this citedness varies between subjects and in this case, is dependent on whether the articles are OA or TA. Measuring the mean citation advantage of OA articles and attempting to find a correlation or association with other bibliographic features helps identify possible causation and links.

8.3.1. Citation advantage

OA articles receive more citations than TA articles, for all subjects from both rounds of data, irrespective of the type of sample taken, whether purposive, randomly chosen or a mid-range sample. This result is in line with previous research..

In comparison to the results here, Antelman (2004) found a citation advantage of 91% for mathematics; the result here was 53%, climbing to 71% when self-citations were removed. As discussed earlier, Antelman took pure maths as her subject. The arXiv

repository holds a large number of pure maths papers, which has a stronger preprint culture than applied maths. Despite this, the comparable result is considered a good match, as, it is suggested, there is only a limited propensity to use arXiv and Citebase to self-archive applied maths articles.

In contrast to these results, Davis and Fromerth (2007) looked at articles from four maths journals and found only a 35% citation advantage for those articles that had been archived in arXiv as opposed to those that had not. Davis and Fromerth (2007, p.205) did, however, use MathSciNet to count citations rather than the ISI databases. MathSciNet does routinely contain self and fairly unusually, preprint citations as well. A small trial of the two databases shows them to have differing levels of coverage and hence, not unexpectedly, different results. The level of uncitedness reported by the two authors is notable at 32.8% unlike the 15.8% here. So, although the results differ in their citation advantage, this is more likely to be a feature of the database used to count citations and the choice of journals and their relative impact factors.

Hajjem, Harnad and Gingras (n.d.) found a citation advantage for both sociology and economics. For sociology, they found an advantage of 172% and for economics, 49%; the comparable results here were 103% and 77% respectively for first round data and 94% and 82% for the second. All the results excluded self-citations. The result for sociology at 172% looks very high in comparison and no reasonable explanation can be given for this, although the result for economics looks more plausible given that the sample taken here was from mid-to-high impact journals.

An interesting result is the increase in citation impact (excluding self-citations) between first and second round data for the three subjects. The result for sociology is marginally in the other direction from 103% to 94%. This is perhaps a reasonable result given the sample from 2004 is a year younger than the original 2003 sample, and has less time to accrue citations. The rate for economics moves from 77% to 82% and for ecology from 49% to 135%. For economics, the difference is not great at only five percentage points. What appears remarkable is the difference in the result for ecology. There are clear differences in the sample taken for ecology with a purposive sample for the first result and a completely random sample for the second. However, it is noticeable that TA articles have a greater percentage of non-cited articles at 31% than OA articles at 19%.

Even when these results are removed from the calculation for those articles, which appear in journals, which have an impact factor greater than four; effectively removing the influence of any possible outliers, the results are still sharply in contrast with the purposive sample. The recalculated gross citation OA advantage is 72% and when net of self-citations, the result is 102%. These results suggest that an OA advantage is evident almost throughout the range of journals.

8.3.2. Citation distribution

The rates of citation for the subjects here vary with mean citation counts being highest for ecology and lowest for applied maths. Characteristically, citation distributions are skewed positively when per article citation counts are made and in any sample of articles, the rate of non-citation varies. Within each subject and in both rounds of data, TA articles always had a greater percentage of un-cited articles than their OA counterparts did. Figure 6.3 shows, for the first round data for all subjects the marked difference in zero citation counts between OA and TA articles. TA articles have a higher level of non-cited articles and 51% of its citation counts fall between one and five, unlike OA articles for which the comparable percentage is 41.7. This difference is particularly evident for sociology, which has the highest level of uncitedness. A distorting factor is its low level of OA articles. Ecology, which has the highest level of citation, has a noticeably different citation distribution (as shown in Figure 6.4) with far fewer zero count articles and has the highest standard deviation. The characteristics of the ecology sample was however, markedly different, needing only seven journals to make up the 1171 articles compared to 16 for applied maths, 22 for economics and 19 for sociology. Further differences emerge in the distribution of citations. For economics, in every citation count between 0-21 there are more OA than TA articles; for applied maths, this is true for citation counts between 1-26. For ecology, by contrast, with one exception, there are more TA than OA articles in the range, for counts between 0-14, and for sociology there are a greater number of TA articles in the range of citation counts between 1-18.

This result reflects the fact that economics and applied maths have the highest level of OA articles. The relative OA advantages are then dependent on different types of citation distribution with generally greater ranges, median values and interquartile

ranges as shown in the boxplot at Figure 6.6. The OA advantage to a certain extent is dependent on the greater range of these citations and their more skewed distributions. In every case, the distribution of OA citations is broader with greater standard deviations and median values. A Chi-square test showed that it is statistically more likely that an article will be OA the higher its citation count, a finding which Davis and Fromerth (2007) also found. The level at which this association is apparent drops to five citations or more when only other author citations are counted. Conversely, of course, those articles with fewer than five or six are more likely to be TA.

8.3.3. Self citations

An important factor in citation analysis is the rate at which authors cite themselves. This can affect the results or conclusions that may be drawn. The rate of self-citation varies within the subjects chosen here but does not affect the overall OA advantage. The actual frequency with which authors cite themselves is highest for mathematics, where 40-50% of all citations are some form of self-citation, compared to about 20-30% for the other subjects. In every case, OA articles had a greater overall mean rate of self-citation, although the differences in mean rates are relatively small. There were also a greater number of TA articles that did not have any self-citations. When examined at subject level, there is a striking difference in the distribution of ecology self-citations, which is the highest amongst the four subjects with 88.3% of all articles having some form of self-citation. This then seems to follow a declining trend towards sociology, which has 55.80% of its article having some form of self-citation. When broken down into citation type, the percentage rates of author self-citation are consistent, with applied maths having the greater number of actual author self-citations irrespective of OA status and sociology the least.

So it appears that the more science-orientated the subject, the higher the rate of self-citation, and having the highest number of articles containing some form of self-citation has limited the citation advantage of ecology when only other author citations are counted. However, a real difference is evident between first and second round data for ecology, possibly because in the second round, the sample was random and much more broadly based. The mean number of all self-citations for OA articles fell from 4.8 to 3.5 and for TA articles this fall was from 3.6 to 2.2. Similarly, the mid-range second round

economics journals had lesser mean self-citation rates falling for OA articles from 1.4 to 0.8 and for TA articles from 1.2 to 0.5. These differences for ecology and economics seem to suggest a higher rate of self-citation for higher impact journals. This result perhaps suggests that more successful or mature writers have a larger body of work from which to cite from, than early career researchers who are publishing initially in lower impact journals. However, when comparing the second round data for sociology, the results are very close to the first round results, with OA articles still having a greater rate of self-citation than TA articles, and with a high number of all articles having no self-citations at all.

8.3.4. Within journal comparisons

Davis and Fromerth (2007) found that in their sample of mathematics journals that the distribution of citations for articles within individual journals was highly skewed. They found that 12% of the articles received 50% of the citations and 44% of the articles received 90% of the citations. This skewing, they reasoned, was evidence of an inequality in the quality of the articles within their journal sample. This skewing effect was noted by Seglen (1992) who found that 15% of a journal's articles get 50% of the citations and 50% of the articles sampled accounted for 90% of the citations. The results found here were not that dissimilar, with 16% of the articles accounting for 50% of the citations and 55% of the articles accounting for 90% of the citations. The results for both OA and TA journals remain almost constant, (as shown in Table 6.12 and Table 6.13) and there is very little variation in these rates. The OA and TA articles are skewed more or less equally. So the argument in this case that there might be a quality bias because of a greater number of more highly cited OA articles within each journal is less sustainable. While the data from Davis and Fromerth (2007) also shows that there are more OA articles with a citation count greater than five, the difference is only 2.6% (14.9% OA, 12.3% TA) and the difference here is 1.5% (3.1% OA, 1.6% TA). A difference, it is argued, that is insufficient to show that a quality bias is evident in favour of OA articles based on skewed citations counts, or a marginal preponderance of greater citation counts for OA articles, at least for the data collected here. The same result is apparent at the subject level, where beyond the five citation count, the only difference of note occurs for ecology (5.4% OA, 8.4% TA).

In fact, when each of the journals from the first round data is examined individually as shown in Appendix F, 54 out of 64 journals have a citation advantage in favour of OA articles. For the second round data, 22 out of the 27 sociology journals had a citation advantage in favour of OA articles, and for the economics mid-range sample, 19 out 21 of the articles had an OA advantage.

8.4. Correlation, associations and OA status

There is a correlation between other author citations and total self-citations, for both OA and TA articles. Attempting to find correlations that would lead to identifying causal factors for OA status effectively failed. Comparing the number of authors and total citations or comparing impact factor and the number of authors all produced very mixed and inconclusive results.

However, some associations that have been discussed above and also some of which appear were related to the frequency of OA status for four variables, i.e., authorship levels, citation frequency, impact factor and country of origin. For these four variables in the first round data there was a modest to good association with OA status, but there is some variability depending largely on the subject. In the second round, the results are more variable with a good association for higher impact factors and the OA status of sociology articles. This was not the case for economics where the subject was only, however, associated with the number of authors. These results lead to being able to point to some clear relationships. For example, if a sociology article were single authored and its author was affiliated in the UK and it was not heavily cited, there is a strong probability that the article would be TA. Similarly, an economics article written by a senior author affiliated in North America with at least one other author has a high probability of being OA rather than TA. Whilst this does not explain why OA articles have a greater citation advantage, it is evidence that there are factors in these articles that help point toward identifying their OA status.

8.4.1. Authorship levels, subjects and OA status

The analysis of both first and second round data allows for the examination of authorship levels and the frequency at which OA articles occur. Table 6.4 clearly shows that OA articles for all four subjects have a greater mean number of authors than TA

articles and that single authored articles (as shown in Table 6.5), are much more in evidence for TA than OA articles. Eysenbach (2006, p.693) also found when he looked at 212 ‘gold’ OA articles from a sample of 1492 published in *Proceedings of the National Academy of Sciences* that they had a greater number of authors than the TA articles.

It is only at the single author level that there is a greater number of TA than OA articles. Chi-square tests confirm that there is an association between this for the results as a whole, but this breaks down at subject level where the association is not significant for ecology and sociology. The almost even split for ecology and the very strong dominance of TA status for articles from sociology irrespective of author count is interesting. Whilst the association at subject level and OA status is variable, the Chi-square test for association between OA or TA status, the author count and subject is strong. This helps confirm, for example, the result for sociology where single authorship is so common. It is only in economics that there are a greater number of OA articles within each author count, perhaps reflecting a preprint culture and the availability of RePEc as a recognisable repository. Similarly, although less strongly evident, the result for applied mathematics shows that there are only marginally more TA than OA single authored articles, although as mentioned later a self-archiving culture is less evident.

In the second round data for sociology, the results were very similar with all authorship levels being dominated by TA articles with almost exactly the same percentage of single authored articles. This level of authorship is typical of the social sciences, where a single scholar approach to doing research is evident and is in contrast to the more team-oriented approach of the sciences where multiple authorship is common (Nederhof 2006, p.88). The result for the random sample of ecology articles was very similar to the first round with OA articles having a marginally greater number of authors, despite there being a much higher percentage of TA articles in the sample, with every authorship count having more TA than OA articles. Despite the randomness of the sample for ecology, the general consistency in authorship levels for this subject and sociology suggests that there is an established structure and culture of shared or single authorship within these subjects. This is not the case for economics, where the number of single authored articles rises from 32% to 41% for the second round articles but

drops from 44% to 40% for those articles with two authors. It may follow that the lower the impact factor, the more single authored articles there are, and the greater frequency of TA articles given the evidence from above that single authorship is linked with TA status.

8.4.2. Country of origin

These basic author counts and OA status characteristics start to diverge when articles are broken down by first author affiliation into their country of origin. In North America and Europe, single authorship is associated with TA articles and thereafter OA for other author counts, but in the UK and the Rest of the World, TA status dominates in every authorship category. North America dominates in its share of the article count, accounting for 57% of the articles, perhaps not surprising given that the first round data was taken from high impact journals. What is evident is the overwhelming TA effect sociology has in every region. If separated and the subjects are examined just by region and subject (as shown in Figure 6.20), a different pattern emerges. Apart for sociology, in which every region is TA, North America has a greater number of OA articles within each subject and the same is true for Europe with the exception of ecology. With the exception of The Rest of the World, applied maths has a greater number of OA than TA articles, with a similar result for economics but with the UK having fractionally less OA than TA articles.

The two rounds of data allow for comparisons. Taking the second round data for sociology shows a weakening in the percentage of TA articles. The percentage share by first author affiliation from North America rises from 57% to 70% at the expense of the UK and continental Europe. The RoW holds its own at around 8%. However, this share of the total article count for North America does not fully account for all of the increase in OA articles; rather this goes to Europe and the UK, with just an additional one percent to North America. As discussed earlier, this may be attributed to a greater awareness of self-archiving and the growing number of institutional repositories in the UK and Europe. The mid-impact second round economics data shows the waning influence of North America, with its share of the article count dropping nine percentage points to 48%. The difference is taken up largely by Europe and the RoW. A much larger difference is evident for the second round random sample of articles for ecology,

where the North American share of the articles drops to 39% from 53%, with Europe and the RoW picking up the difference. Evident also, although on a smaller scale is the general decline of the UK's share of the articles the further down the journal impact factor scale, the journal appears.

This presence of North America in terms of authorship and level of OA is also reflected in the work of Hajjem, Harnad & Gingras (n.d.), and of Eysenbach (2006). Eysenbach showed that in his overall article sample, the United States had the majority of articles at 65.8%. However, it was only ranked fifth in the level of OA articles in the sample, but it should be remembered that this 'gold' sample of articles was self-selecting by its authors. Interestingly, the UK was amongst those countries that had the lowest level of gold OA articles in the sample, lending some small support to the general preponderance of articles to TA in the Eysenbach sample above.

8.4.3. Impact factor and OA

Authors from North America appear more frequently in the higher impact journals when compared to their appearance in the mid-range sample for economics and the random sample from ecology, affecting, it seems, the rate of OA. An example of this is apparent when comparisons are made between first and second round data for economics and ecology. The level of North American authorship drops noticeably. The OA rates change for these subjects from 53% to 34% in the second round for ecology and for economics, this drops from 65% to 54%. First round data indicates that there is a reasonable association between OA status and IF, with boxplots (shown in Figure 6.25 and Figure 6.26) confirming this result. Similarly, second round results for sociology also indicate a strong association.

Results from the second round sample for economics were taken from mid-range journals, where the range of impact factors was very narrow, from 0.398 to 0.967 as opposed to a wider range for the first round data. There was no association between IF and OA status, suggesting that the narrow IF range may have limited the result. The second round ecology data, however, provided a better opportunity to test this assumption, given that the sample was random and articles were taken from journals whose impact factors spanned the entire range within the ecology subject category. In

comparison to the first round data, 84% of the journal impact factors lay outside those taken from the first round. Coincidentally only 16% of the articles were drawn from the same impact factor journals as those sampled in the first round ecology data. However, they did make up, not surprisingly given their high impact status, 30% of the articles. The results shown in Figure 7.37 show a clear inter quartile difference and median value for the IF in favour of OA articles. The bar chart in Figure 7.38 strongly suggests, given the negative skew for the OA articles and the relatively random distribution for the TA articles, that the higher the impact factor, the more likely an article will be OA.

Suber (2005d) appears to confirm this result, because in his comments on the work of Wren (2005), he suggests that authors who publish in high impact journals are more willing than those who publish in lower impact journals to make their articles OA. Suber checked the journals that Wren (2005) had used and found that lower impact journals were less likely to permit self archiving, so giving a plausible reason why higher impact journal articles are more likely to have been made OA. However, Antelman (2006b) could not find any relationship between publisher policy and the self-archiving practices of authors. This suggests that authors were self-archiving their articles irrespective of any particular publisher policy. Her work was related to the social sciences rather than the biomedical journals that Wren examined.

8.5. Discovery of OA articles.

Four different search tools were used to identify the OA status of articles with varying degrees of success. Harnad *et al*. (2004, 2005 & 2006) have long argued that authors wishing to make their articles accessible to others should self-archive them to institutional or subject repositories on the basis that they can be both readily found and cited by anyone searching for them on the World Wide Web.

The 2280 OA records in the first round of data collection were found using the four different search tools as described in section 6.7. *Google* and *Google Scholar* were far better at finding OA articles than OAIster or OpenDOAR. When examined overall and at regional level, *Google Scholar* far outstripped the other search tools in every subject. When the OA articles were broken down by first author affiliation and subject, North America was found to be the region from which most articles originated, apart from

applied mathematics, where it accounted for 43.2% of the articles. For the other subjects, the percentage article counts from North America were, ecology 61.9%, economics 70.4% and for sociology 77.5%. Clearly, taken in combination *Google* and *Google Scholar* were far more successful than OAIster and OpenDOAR, whose success was relatively poor. Whilst the UK has a relatively modest overall OA record count of 201 articles, it has the highest combined percentage for those found at OAIster and OpenDOAR at 24.9%, suggesting that whilst self-archiving is still relatively poor in the UK, some of those who self-archive are choosing to use repositories.

Only in economics and applied maths could OAIster and OpenDOAR be considered useful search tools, finding 21.2% and 20.4% respectively of the hits between them. Given that both OAIster and OpenDOAR list the economics database RePEc among the sources from which they collect articles records, it is not surprising that there was a reasonable number of hits in this subject when using these two search tools. It is less clear why applied mathematics, whose finds were at a similar level using these search tools, was similarly successful. It is noted, however, that OAIster and OpenDOAR do list arXiv and Citebase among their resources and whilst arXiv is a resource more focussed on pure mathematics, a number of hits were found both there, at Citebase and at institutional repositories. Obviously, if authors use repositories where their records can be harvested by these service providers, then the recall from them will increase.

It is notable that in sociology, which had the fewest OA articles overall, that the majority of them were found using *Google* and *Google Scholar*, suggesting that those who do self-archive their work tend not to use repositories but use home pages or departmental sites. This seems inevitable since subject repositories like arXiv and RePEc are not available to the social sciences and the humanities in general.

Swan and Brown (2005, p.28) surveyed, academics' self-archiving preferences, for a range of disciplines. They found that authors in mathematics were roughly as likely to self-archive on a web page as a repository. Similarly, for those in the social sciences and education it appeared just as likely that they will self-archive to a homepage as a repository. The sample from their survey was relatively small and self-selecting, and so their results, perhaps not surprisingly, do not reflect the results found here.

When the second round data is compared to the first round data, there are some notable trends. For sociology, the percentage of hits drops for *Google Scholar* but rises for *Google*. Overall, their combined share of hits drops from 98.37% to 93.38% to the benefit of OAIster. This is in contrast to the hits in ecology, where the combined *Google* score is 96.27% for the first round but for the second round, this rises to 99.10%. Given that institutional repositories are more likely to be found at the more successful institutions (Directory of World Repositories 2008) and that the sample for the second round ecology data was randomly taken and hence more likely to come from a range of different institutions, it could be argued that it is more likely that the authors would self-archive to their own websites if repositories were not available at their own institution. However, for economics, where OAIster was particularly successful, the combined scores for *Google* drops from 78.76% to 60.68% giving 39.32% of the share of the hits to OAIster and OpenDOAR in the second round, perhaps mirroring the growing success of these harvesters.

Whilst *Google Scholar* is obviously the most successful, there are some successes for OAIster and OpenDOAR. When these first round differences are examined regionally, there are some notable features. Outside of North America, the success for OAIster and OpenDOAR comes principally for economics and applied maths. The Rest of the World leads in economics and the UK leads in applied maths and ecology although the article count for both of these is relatively small. When second round data is examined, the trend for the three subjects, except for ecology is towards OAIster and OpenDOAR having greater coverage. This coverage when examined at regional level for economics shows that Europe leads with the greater percentage of hits with a general fall in the reliance on *Google* and *Google Scholar* across all regions. Comparing the results from Table 6.9 for the first round economics data where 78.8% of the hits were from *Google* and *Google Scholar* compared with the results shown in Table 7.6 for the second round shows this advantage dropping to 60.7%.

The relative success of OAIster and OpenDOAR is attributed to their harvesting the metadata from RePEc and the need to share informal research results in general symposia and in working paper series (Antelman 2006, p.92). Antelman (2006b, p.89) examined self-archiving practices within the social sciences, taking approximately 2000

articles from 22 high impact journals from 11 different publishers with varying self-archiving policies, including economics and sociology. For economics, she found an overall rate of self-archiving of 59% and for sociology 24%. For the two samples taken here, the rate for the economics' first round data was 65% and for the second round data 53.9% and for sociology 21% and 24.38% respectively, a noticeably similar result. Antelman goes on to explain the overall level of self-archiving as characteristic of the discipline, for the social sciences this is one where authors are less reliant on a culture of sharing information for example in the exchange of preprints. Economics, however, is characterised as a discipline with a higher degree of mutual dependence where working papers are shared through repositories with other authors. Apart from the RePEc there are few disciplinary depositories for the social sciences. Hence, there is little difference between the results between the first and second round data for sociology, with the OA hits being found almost exclusively by *Google* and *Google Scholar* and with few academics archiving to any sort of repository.

Despite the increasing number of institutional repositories (Registry of Open Access Repositories 2008) and their harvesting by such services as OAIster, it is apparent that finding OA articles in the four subjects selected here was greatly facilitated by the use of *Google* and *Google Scholar*. What is clear that whilst OAIster and OpenDOAR are reliant for the majority of their content from institutional and subject repositories, it appears that the majority of authors in this sample at least are not self-archiving their work to them or if they do, it is to non-compliant or to unregistered repositories. Authors prefer, it seems, when they do self-archive their work, to do so to their personal or departmental web page where metadata harvesters such as OAIster cannot readily find them, but where *Google* and *Google Scholar* can. However, it is clear that both OAIster and OpenDOAR are starting to increase their coverage, due in part at least, to the increasing number of repositories and the drive to get academics to self-archive their work.

Bergstrom and Lavaty (2007) used *Google, Google Scholar* and OAIster to help them find OA articles in economics and political science. From a sample of 703 economics articles, they could find most OA articles using *Google*, with *Google Scholar* finding some ten-percentage points less than *Google*. They found, using OAIster about 25% of

their sample articles. This is a similar result to those found here, where 21.2% of the articles were located by searching OAIster. RePEc provided 27% of the articles, which is in itself an interesting result given that OAIster lists RePEc as one of the sources it trawls. When the holdings of the two sources are compared, it is clear that not all the records available from RePEc are reported by OAIster and presumably, this explains the difference, although it is very unlikely that there were any items discovered in RePEc that could not found by using *Google* or *Google Scholar*.

8.6. Causation – in general.

Since the work by Lawrence (2001), the causes of this advantage have been much discussed. Initial work by Anderson (2001), Harnad *et al*. (2004, 2005 & 2006) and Antelman (2004) was unable to isolate the causes of the OA citation advantage, although Harnad uses the citation advantage to support his drive for self-archiving amongst academics. If it could be shown that making an article OA would ensure an important citation advantage, then this would be a powerful argument to self-archive. Kurtz and Henneken (2007, n.p.) suggest there is no citation advantage, which can be attributed to the fact that an article is OA, but they think there are many "excellent arguments in favour of changing the scientific publication system to an open access model. The open access citation advantage is not one of them."

An important feature of citation advantage work using the arXiv database is that the articles deposited there are date stamped upon receipt, thereby giving a fixed point from which to calculate and compare any citation advantage. The work using other sources lack this precise deposit record. However, most research has not been able to control for authors depositing their work elsewhere, nor have those authors tried to ascertain that those authors who have not deposited their work in the arXiv have not self-archived elsewhere. Davis and Fromerth (2007), Moed (2007), Kurtz and Henneken (2007) examined the causes of any citation advantage for those articles which were OA. Their results have shown variously that they cannot discern a citation advantage based simply on the fact an article has been made freely available. Rather they suggest that making an article freely available prior to publication and usually archived in the arXiv repository can give a citation advantage. Moed (2007) specifically was not able to identify an open access advantage but rather found a strong early view and quality bias, as discussed in

section 4.18, to explain any citation advantage. Kurtz and Henneken (2007) argued that there was no evidence to support any citation advantage derived solely from making an article open access.

8.6.1. Possible causation: ecology

Second round ecology data was used to see if those articles that were OA exhibited chronological chronological citation characteristics that might indicate a plausible citation advantage based on based on OA status alone. The article records were collected randomly and as well as recording recording their citation counts, the year in which the citations were made was also taken. The series taken. The series of graphs shown in section 7.20 show the profile of the OA and TA article records article records by their mean citation count. The first of the graphs shown in Figure 7.39 shows the shows the overall OA/TA difference and the remaining four show the result by territory. In every In every case, there is a clear citation advantage for OA articles, starting in most cases in the first in the first year of publication and rising steadily throughout. As mentioned earlier, the sample sample ranges across all impact factors for the JCR subject. As the further graphs in

Figure 7.40 illustrate, even when those journals that account for 30% of the citations from the higher impact journals are removed there is still a noticeable OA advantage for all articles across all territories. When this result is taken into account and the almost universal finding for first round data and that of sociology and economics from the second that within each journal there is a citation advantage for OA articles, it would seem that there might be some contribution to this advantage from the fact that these articles are OA.

8.6.2. Possible causation: OA and poorer countries

Whilst there is evidence from the data collected here that there is a citation advantage to those articles that are made OA, the actual causes of this advantage are unclear. One of the primary arguments in favour of OA is that those who cannot afford access to peer reviewed journal articles could do so if the authors of these articles self-archived their work on the World Wide Web where they could be readily accessed. It should follow then that a higher percentage of those who cite these OA articles ought to come from countries where access to expensive journals is limited. Hence, it could be reasoned that

if this were clearly so, that a demonstrable cause of any citation advantage could be shown. Smith (2007) posed this question in a contribution to a discussion list. On the other hand, it may be that the lack of reliable telecommunications networks in these countries could hinder access to OA articles. In this case, scholars in these countries may rely on a limited number of printed journals for which they have subscriptions. Notwithstanding this the analysis in section 7.32 examines, for applied maths, the country of origin of articles by their first author affiliation and the origin of those authors, which cite them, again by first author affiliation for any evidence which might help identify the causes of any citation advantage.

Table 7.12 shows by country of first author affiliation the distribution of cited and citing articles by their OA status and their classification by the per capita income of the affiliated author's country. For OA and TA articles, the highest ratio of citing to cited articles occurs for those countries in the lower middle income bracket. If all but the high-income level of articles are taken together, then there is a twofold increase in the citing ratio of OA articles when compared to the TA articles. Whilst this appears to be a convincing advantage the actual percentage of lower income TA citing articles, is 20.0% and is greater than the 15.4% for the comparable calculation for OA citing articles. So there is a greater percentage of lower income TA citing authors than OA citing authors. The reverse would have been expected given that the results from section 6.8 suggest that overall articles from high impact journals are more likely to be OA, and if a quality bias were evident then it would be expected that better articles are self-archived and hence more likely to be cited.

Given that self-citations have been eliminated in this analysis, it is authors citing the work of others to support their work, who are doing the citing. As can be seen in Table 7.13, however, most of those who are doing the citing to the lower income groups are from the high-income countries. It is clear that a greater percentage of authors from the low and lower income countries cite more TA articles than OA articles, despite there being a higher ratio of citing to cited articles for those of OA status. In fact, 95.3% of all OA citations and 86.5% of TA citations are from high income regions. There is little evidence to support, for this subject and these articles, that making articles OA aids access and hence a greater rate of citation by those who may not be able to afford access

to expensive TA journals. It may be simply that aspiring countries and scholars find it essential to access high impact journals at the time of publication rather than relying on authors routinely self-archiving their work. As one of the central tenets of the OA movement is to make scholarly peer reviewed articles freely available on the Internet for anyone to cite and read; these results are clearly disappointing.

A broader picture of the distribution of citations by OA status and their origin by first author affiliation is shown in Table 7.14, where they are matched to the country of origin of the articles they are citing. Generally, for every citation that can be paired by country to the article it is citing, there are three that do not match. This applies to both OA and TA articles. Clearly, however, this data is skewed by the predominance of the region-to-region match for the USA. Given that just over 25% of the citations come from this territory alone, it is not surprising that of all citations almost 42% are to USA affiliated authors. Perhaps this result is unremarkable given that most (38.6%) of the cited articles originate from the USA and that of all the citations from each region that the majority are to USA affiliated authors. This picture of citing ratios emerges (from Table 7.15 and Table 7.16) where cross tabulations show the distribution of all citations to cited articles by country of origin. Of the 230 (90 TA, 140 OA) citations made by Chinese authors this represents, of the total 3032 citations (935 TA, 2097OA) 9.6% of all the TA articles and 6.7% of all the OA articles. Similarly, for the Pacific Rim and the Rest of the World territories, the TA/OA citation percentages were barely different at around 5% each. This result appears to confirm the findings from the analysis by per capita income, that is, that there is little evidence to suggest that authors who live in countries that may have difficulties accessing TA journals are citing OA articles in greater numbers and hence boosting the citation count. Figure 7.60 shows that seven out of the twelve regions have more citations to TA than to OA articles. The seven regions include the UK, Italy, Japan, Spain, China and the Rest of the World, the latter two helping to support the premise that low-income does not generate exceptional OA citations. It is noticeable also, as demonstrated by Table 7.14, that the regional link between cited and citing article by first author affiliation is weak. For citations of either OA status, the overall regional match is about a quarter, but noticeably in the case of China, this is heavily skewed in favour of not citing other Chinese affiliated authors.

8.6.3. Causation in collective subjects

Logistic regression was used by Eysenbach (2006) in his work on open access advantage to identify factors, which might be distorting the OA advantage that he found. Logistic regression has been used here in Objective 5 of this work to try to identify, variables that could help determine the observed frequency of OA article. This is an entirely different approach to that used by the authors reported in section 8.6, who were concerned with identifying the cause of any citation advantage that might be derived from a self-selection bias, an early access advantage or the quality of the article. In this case, the advantage of having articles date stamped at the point of self-archiving becomes less of an issue.

The results from the logistic analysis for all 4633 article records showed that it was initially possible to predict whether an article was OA or TA simply by counting how often any article appeared in either group. In this case, it is marginally more likely that an article will be TA than OA, given that TA articles make up 50.8% of all articles. After analysis, the model showed that it could improve the rate at which it could predict the level of open access to 67.2%. Out of the 18 initial independent variables entered into the model, 10 were discarded as being not significant ($p>0.05$) and hence having no effect on the ability of the model to predict the rate of OA, leaving as the main predictors the subject categories, country of origin, number of authors, impact factor and the number of other author journal self-citations.

These main predictors varied in their influence, the subject categories of ecology and sociology had the greatest effect on the model in deciding whether an article was OA or TA, in the case of these two, the influence was towards predicting that an article was TA. This is not surprising given that so few of the sociology articles were OA and just over 50% of the ecology articles were OA. However, those variables which did have a positive influence in helping predict the OA status of an article were the countries of origin by first author affiliation from the USA and Europe and the total number of citations that an article received. The rate of influence of any of these variables was not high, inasmuch that if they were to increase, their odds of increasing the likelihood of being able to predict whether an article was OA or not was quite poor.

8.6.4. Causation in individual subjects

When individual subjects are subjected to analysis using logistic regression, clearly individual subject status is removed as a potential predictor variable. The results for the first round subjects show only marginal improvements in the model's predictive power. Notably in three of the subjects, the level of citation plays an important part, albeit small, in the change in the models predictive power. For sociology the model predicts more readily the TA status of an article than it does when all the subjects were considered together. Only in ecology can a significant result be shown. This is due, almost certainly, to the dominance of the USA by first authorship affiliation with 52.86% of ecology articles originating from this territory. A similar result is evident in the second round results as shown in Table 7.26 for sociology and ecology but the results for maths and economics are inconclusive. The ecology article records were selected randomly and it is not surprising that the dominant predictor of OA status is impact factor given that mean impact factor for OA articles was noticeably higher than the that for TA articles as shown in Figure 7.37.

In summary, analysis by logistic regression indicates that the largest factors in deciding the OA status of an article is its subject (and by extension that subject's culture) rather than, say, the number of authors, impact factor, self-citation count, etc. Too much of the variability is unexplained by the model and what is explained is dominated by the subject. Much like astronomy and physics and the propensity to self-archive to arXiv, it is almost axiomatic that the majority of these articles will be OA at sometime. Looking at the subjects individually is even more difficult with no one variable standing out as explaining the variance, added to which there is only a marginal increase in correct category prediction.

Nevertheless, a modest part of the variability in being OA can be explained by the subject. In the case of sociology, the indications are that this effect is negative and with reduced odds, as you would expect given that only 20% of the sample was OA. Similarly, economics and applied maths have an impact on the status of OA articles and this again is borne out by the data, being the most OA subjects. Supporting this is the positive influence of the USA helping in the prediction of OA status. Again, this is not surprising given that the USA generated the most articles and has the highest incidence

of OA articles. What is not evident is that author count, self-citations, other author citations and impact factor have formed a substantial part of the variance independently of any of the subjects. This means that it is not possible to point to any particular set of characteristics that might explain why an article is OA in the way that say the early access postulate could be as suggested by Kurtz and Henneken (2007). So in essence, casual interpretation from the statistical analysis rather than logistic regression is a good guide to gauging the likelihood of an article being OA. This seems hardly surprising given the difficulty that authors have had in isolating what might be the cause of any OA advantage, except those of course who have inched their way towards some conclusions for those articles deposited in arXiv.

8.7. Conclusion

The OA citation advantage was apparent for all four subjects, for both rounds of data and in the majority of cases, at journal level as well. This citation advantage varies, however, between subjects with sociology having the highest citation advantage. Perhaps the most interesting result is that from the second round random sample for ecology, where the citation advantage not only rises but also is even greater when high impact journal counts are removed. The data also indicates a pervasive degree of OA articles across the range of impact factors. The notion that self-citation might boost any citation advantage was dispelled when the results demonstrated that without them, the citation OA advantage was increased. Economics and applied maths exhibited higher levels of OA and this was attributed to the availability of disciplinary archives such as RePEc and arXiv where authors were archiving their work as a matter of course and a growing culture perhaps, where authors feel they need to disseminate their work amongst their peers. This, it seems, is a more noticeable feature of those authors publishing in the higher impact journals, the higher the impact of a journal, from the data here at least, the more likely it will be OA.

Searching for OA articles was facilitated by using different search tools and the success of these in finding articles indicated two things, their efficiency in finding articles and hence their coverage and as a corollary, where self-archiving authors were placing their work. *Google* and *Google Scholar* were the most successful search tools. OpenDOAR and OAIster were limited by only returning hits from compliant repositories, and

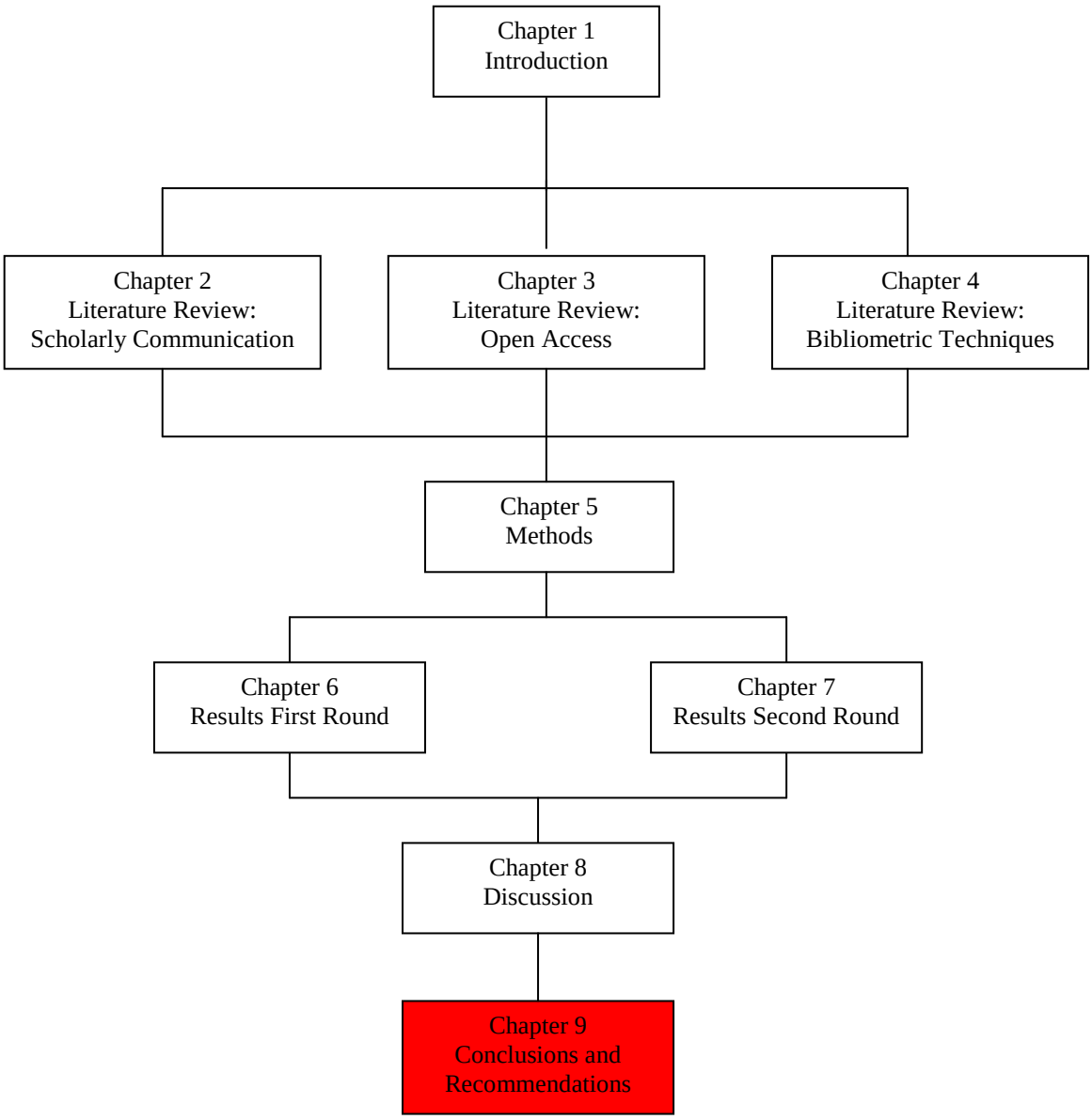
naturally were more successful in finding economics and applied maths articles. Perhaps more interestingly was the overwhelming use by academics to self-archive to personal or departmental web page rather than institutional repositories, indicating a possible lack of these, or that authors simply found it more convenient to use their own resources.

North America dominated, in terms of the number articles by first affiliated authors and the level of OA in all subjects. This domination waned the lower the impact of the journals from which the articles records were drawn and the lower the rate of OA for the mid-range economics and the ecology random sample. Where North American dominance receded, the advantages spread to Europe and the RoW. Although not as marked in the UK there was a tendency for the share of articles to drop the lower the impact factor of the journal.

The number of authors an article has seems to be one factor in deciding whether an article is OA or not, those articles with two or more authors are more likely to be OA. The single scholar approach evident in sociology links in well with the high level of TA in this subject although unlike other subjects there are no well-defined subject repositories where authors can self-archive their work. Where first author affiliation is from North America, they are more likely to self-archive their work than those in the UK and the RoW. Similar associations are evident in the second round of data, but an interesting difference was evident for the mid-range sample of economics articles, where single authorship and the level of TA rose.

Looking for causal links between the OA status of an article and its bibliographic details proved to be inconclusive, being no better than just perusing the results obtained from the basic data. There was also no association between citations to OA articles with authors from less well developed countries where access to costly journals may be difficult. What was evident from the random sample taken for ecology was a persistent OA citation advantage even if high impact journals were removed. Overall, it seems that given the persistence of the advantage across the subjects and in both rounds of data that some contribution from the OA status of an article to the citation advantage is plausible.

Chapter 9 Conclusion and Recommendations



9.1. Introduction

Peer reviewed articles that are made open access allows any reader who has access to the Internet the opportunity to read and cite the work, quite apart from increasing any research impact the work may have from being accessible. It seems plausible then, that if an article is made freely available by their authors, that it will be read and cited more often than those that are not. This final chapter considers the consequences of this assumption in relation to the foregoing work. The contribution and the implications of the work are considered as well as any limitations. Finally, a number of recommendations are made as well as suggestions for further research.

9.2. Main findings

There was a clear citation advantage for those articles that were OA as opposed to those that were not. This result was evident for all of the four subjects irrespective of where the sample articles were taken from and in almost every case, this was also evident at journal level. While it can be clearly stated that this citation advantage is unmistakeable, it is evident that the levels of OA varied amongst the subjects, indicating it is believed, different academic cultures where in some subjects self-archiving is readily undertaken and is facilitated by suitable subject archives.

The causes of this advantage were less clear; the results were not strong enough to point convincingly at definite causal links. This finding was also in keeping with the results that failed to show that poorer countries were accessing OA articles and citing them more frequently than TA articles.

9.3. Contribution of the study

Earlier studies, with the exception of Anderson (1999) showed that there was a consistent citation advantage for those articles that were OA. The majority of these studies took their samples either from arXiv, where a very strong preprint culture exists, or from a broad range of subjects. Antelman took a sample of four subjects and Eysenbach (2006) looked at articles in a range of science subjects that were made open access by payment.

9.3.1. Citation advantage

In line with Antelman's (2004) work, this study focussed in depth on four subjects. Applied maths and ecology had not been examined before. Unlike other studies, a second sample confirmatory sample was taken for sociology and economics from 2003 to check results, and showed the advantage to be persistent. It was clearly demonstrated in the study that self-citation did not affect any citation advantage.

9.3.2. Levels and distribution of OA articles

Subjects vary in their level of OA and in so doing indicate where this may be an issue, particularly where exposing the collective research effort of an institution is important. Almost uniquely, with the exception Gingras *et al.* (n.d.), the level of OA was taken a step further by examining this at a regional level and this showed for example that the UK is the poor relation to North America and marginally to Europe as well, but arguably the UK is catching up.

9.3.3. OA article characteristics

The study showed that OA articles exhibited certain characteristics. For example, the number of authors an article had could be correlated with OA status. This characteristic was, however, regional with North America and Europe the strongest. Similarly, it was evident that the subjects exhibited strong authorship cultures; sociology was strongly associated with TA status and single authorship. It was clearly shown that the higher the impact factor of a journal the more likely that its content would be OA; the random sample for ecology demonstrated this particularly well.

9.3.4. Search tool success

None of the earlier work that used search tools to find OA articles measured their success. Knowing how successful these search tools were at finding OA articles would suggest how best to guide authors in their search, or where deficiencies in the other search tools might be addressed. The success of *Google* and *Google Scholar* was almost complete when compared to OpenDOAR and OAIster but it was apparent that these latter two had strengths in finding applied maths and economics articles.

The findings suggest those wanting to find OA articles, would be well advised, to use *Google* or *Google Scholar* rather than OpenDOAR or OAIster, until their coverage

approaches that of the *Google* search engines. It appears, *Google* and *Google Scholar* are accessing the same institutional repositories as OAIster, and are gaining in addition access to homepages and departmental websites from which to draw academic material; it seems likely they will always have a greater recall even if the structure and format of the returned hits are less formally presented.

9.3.5. Causation

The causes of any citation advantage are difficult to determine. Kurtz and Henneken (2007) are confident that they can show for astrophysics that it is a feature of early access. Harnad (2007a) has speculated that he thinks that there is at least a combination of early access, quality bias and some small element of open access in any citation advantage. Outside the culture of the astrophysics and arXiv community, it still seems possible that in other disciplines there may be some citation advantage which can be attributed to the OA status of an article. The findings here did show that certain bibliographic characteristics of an article, whilst a useful indicator of OA status, were insufficient on their own to predict this status. The result suggests that using the characteristics of an article to determine OA status is unlikely to be a useful avenue of research. In a similar result, looking at the origin of citing authors by their country was inconclusive, but this part of the study was indicative only and is worthy of further examination. In the longer term it may be that as academics are required to self-archive their work, either by mandate or the need to validate citation metrics used in performance measurement, that any citation advantage, from whatever cause will simply disappears as articles are routinely self-archived.

9.4. Implications

Whilst the causes of any OA citation advantage are debatable, what is evident is that those articles that have been made OA by their authors, in this work and earlier studies, have collectively an almost universal citation advantage over comparable TA articles. Studies in general show that there is a good correlation between citation impact and the academic success of individuals and institutions. So it follows that enhancing citation impact ought to be a priority for individual academics at the institutions where they work and those who fund them directly or indirectly. Such an approach, potentially at least, may be seen to have a broad effect not just on academics and their institutions but

also on the wider network of organisations which support them. Current levels of article self-archiving are estimated at 15%; clearly, this leaves plenty of scope for additional records to be archived and the possibility that further citation advantage may accrue to self-archiving authors.

9.4.1. Government

Government in the UK for example is the prime source of income for universities, either directly or through its funding agencies such as research councils. The UK government has used the Research Assessment Exercise (RAE) to apportion research funding at UK universities based on the peer-review of scholarly output. After 2008, funding decisions will be based more broadly, and are likely to include metrics that use citation counts in some form. Establishing metrics that are acceptable to academics will be challenging. Given that OA articles have greater citation counts than their TA counterparts, then all government-funded research should be made OA, in a timely manner, if bias in favour of greater citation counts for OA articles is to be avoided.

The research has shown that the UK for the four subjects examined has a poor record in making its research freely available. Given that the majority of the work is funded indirectly by taxpayers, it also seems reasonable that they should have unfettered access to its results. World rankings are being developed which not only rank universities but also rank repositories, however distasteful or suspect the process might be. Quite apart from this, as another measure of prestige and standing UK plc should actively promote self-archiving at an institutional level and for example through OpenDOAR the harvesting of their contents. Increasing OA citation impact, however the mechanism works, will potentially help rank individual scholars more highly and their institutions.

9.4.2. Institutions

Increasing the standing of an academic institution brings prestige both locally and on a worldwide stage hence the greater the recognition the greater the likelihood of increased funding. Measures of increased citation impact help, in a small but significant way, to improve the standing of an individual, their department and the institution. Those authors who make their work OA are far more likely to increase their citation impact than those who do not, particularly while the rate of self-archiving remains at its current low level. Institutions should ensure that they have repositories into which academics

can deposit their work, both as a public good and as a platform to show case the institution's scholarship. The current low level of self-archiving should be addressed either by mandating the deposit of scholarly output or, rather, by developing in the longer term a professional culture amongst academics where such activity is seen as routine.

9.4.3. Funders

Notably, those in the USA who receive research funding from the National Institutes of Health have an obligation to share their results by self-archiving them for the benefit of other researchers and for the public who fund them. This is the case for those recipients of Wellcome Trust funding and from most UK government research councils. Given the poor level, for example, of self-archiving in sociology in this study and ironically as it has the highest level of citation advantage, obliging recipients to self-archive their funded work seems to be essential to increasing access and individual citation impact. Such funding bodies will also find it appropriate to monitor self-archiving levels if citation metrics become part of future funding decisions, both to ensure that a minority do not benefit from skewed citation counts and to achieve equitable access to the research. To this end, all funders should mandate the deposit of author postprints to either an institutional or a central repository.

9.4.4. Information suppliers

The *Web of Science* and *Scopus* provide multidisciplinary databases that index a wide range of journals, some of which are open access. In terms of peer reviewed scholarly journals these two suppliers index about half of those that are available. Governments that adopt citation metrics to measure the success of their academic institutions will need, if only to satisfy critics of coverage, to have access to databases, which index a broader range of journals and hence a fairer coverage of citations counts. Such database suppliers might consider accessing the contents of subject repositories or with the help of OAI compliant harvesters the content of institutional repositories to supplement their coverage.

Whilst commercial database suppliers deliver search contents in a highly structured and reliable manner, their citation counts are at odds with those found by *Google Scholar*, whose citation counts are generally higher and more broadly based. Not only this,

Google Scholar is far less conservative in where it finds and returns hits for author searches, including those articles which are OA and their respective citation counts. The results from *Google Scholar* are currently difficult to use; the content cannot be readily downloaded. Services such as those developed by Harzing (Publish or Perish 2008) allow further processing of these records into recognisable and useable metrics. Such processing perhaps gives a more realistic measure of a scholar's citation impact, using the growing battery of performance metrics that can be automatically calculated.

Harvesters such as OAIster harvest metadata from institutional and subject repositories; relatively speaking their coverage is limited. Whilst the way they structure their content is better than general search engines, the rate of recall is inferior to that from *Google* and *Google Scholar*. Given the low rate of self-archiving, it is questionable, for the moment at least, whether these types of services are useful to scholars while ever they are dependent on them to self-archive their work to such repositories. There is a clear argument for *Google Scholar* to harvest metadata from institutional and subject repositories and combine these with its other resources, provided it can deliver de-duplicated search results in a more formal and orderly manner. Attaching meaningful citation counts to these article records would provide a formidable database for scholars and institutions alike. This argument could be just as readily made for *WoS* or *Scopus,* but concentrating on extending their coverage to repositories and beyond.

Publishers are understandably cautious about the long-term effects of the OA movement. It is suggested however, as a long-term strategy, that there is a real opportunity to take the lead in this movement by simply developing access policies to their journals based on the embargo periods they have already agreed through the RoMEO project. For example, Reed Elsevier could allow access to their database for their entire collection based on say a moving six month embargo period. This might be an area where *Scopus* or the *Web of Science* could gain a significant advantage, simply by providing direct links from their databases to OA articles hosted by publishers themselves. If this were a large-scale endeavour, then the need to persuade authors to deposit their work on a mandatory basis to repositories would be removed. The final step would be to have a mechanism that would allow the public access to these databases. If this were not acceptable to *Scopus* or *Web of Science* then perhaps this could be done through *Google*, *Google Scholar* or some industry-funded agency.

9.4.5. Academics

Individuals who make their work open access are shown collectively to receive more citations that those who do not. As a measure of success, citation impact is one metric that can be readily used to identify those academics whose work is recognised and cited by others, including funders, research councils and government. New measures of impact like the *h*-index make it possible for individuals to have their research impact more readily compared with their peers. It follows then that funders and government can use this, and other measures to award funding. The implications for individuals, on this and similar measures at least are that they should maximise their impact by making their work visible to the widest possible audience by making it OA.

9.5. Recommendations

The following recommendations have been derived from the findings of the study and the implications already discussed.

9.5.1. Government

- Should mandate the self-archiving of any research output it funds and ensure research councils monitor deposits.
- Encourage universities and institutions to self-archive scholarly output in an OAI compliant format as a matter of course especially where this has been publicly funded.
- Develop equitable citation and other metrics, which help in part funding decisions.

9.5.2. Institutions

- Should develop policies that ensure publicly funded research is made freely available and accessible to anyone with Internet access.
- Should provide access to their own or shared institutional repositories, or direct authors to disciplinary archives that are OAI compliant.

- Should engage with academics and departments to decide how best to achieve a high level of self-archiving.

- Should explore and develop the use of citation metrics to gain a competitive position amongst their institutional peers, and in the funding market.

9.5.3. General information suppliers

- Widen the coverage of the journals they index if they wish to provide the base data from which citation metrics can be more broadly calculated.

- Investigate the possibility of gaining access to alternative citation indexes to broaden citation coverage.

- Build programs that would allow the ready accurate calculation of citation metrics, so individuals and funders can assess the citation impact of individual scholars, departments and institutions.

- Seek to engage publishers in a joint venture to allow wholesale, routine OA to their journals after a suitable embargo period. This should be through an agency that allows free public access.

9.5.4. Google Scholar

- Allow more sophisticated search parameters to limit hits returned.

- Improve the presentation and accuracy of hits returned.

- Make it possible for users to manage and manipulate the results it gives.

- Broaden coverage, possibly including repositories or by direct access to publisher journals.

9.5.5. Metadata harvesters

- Engage with higher education institutions to develop institutional archives from which they can harvest metadata.

- Encourage those sources from which they do harvest metadata to engage with their academics to self-archive their work more readily to them rather than to personal web pages.

- Consider adding citation metrics to the records they do hold.

9.5.6. Academics

- Engage actively in self-archiving to institutional and subject repositories.

9.6. Limitations of the study and recommendations for further research

Inevitably, there are limitations to the work. Firstly, the scale of the work was limited to four subjects and the sample was in turn limited by the data collected in the first round to high impact journals. The purposive sample taken, limits to a certain extent, how far the work can be generalised, although aspects of the data collected in the second round did confirm many of the initial findings. The scale of the sample was small when compared to the tens of thousands of articles sampled by others.

Identifying OA articles can be problematic. Inevitably, some articles would have been coded as TA when OA versions were available either because they were lost behind broken links or were simply not found because of title mismatches. The study was limited to counting citations from a single database; other citations to articles will have been made from journals, which are not indexed by in this case the *WoS*.

9.7. Recommendations for further research

The cause of the OA citation advantage may have been established by Kurtz and Henneken (2007) for astrophysics articles deposited in the arXiv archive, but it is much less clear and more difficult to prove for other subjects. The rates of OA vary considerably between subjects. For some subjects, it is very high and others, e.g., sociology it is relatively low. Based on the findings of this study the literature to date and the limitations of this study, further research is recommended. Such research should focus on a number of areas:-

1. The study was not able to demonstrate the cause of any citation advantage. Research should be undertaken, which randomly assigns sufficient articles to

OA or TA status at the point of publication, irrespective of the author status or perceived article quality from within a small range of journals and disciplines. Perhaps with the agreement of a publisher such a range of journals within particular disciplines could have their articles randomly assigned to OA or TA status, with the agreement of course, of their authors. The citation counts of these articles could then be tracked to establish whether an OA advantage was evident or not. If such an advantage was evident, this could then be used as an incentive for authors to self-archive and would be an important factor to take into account if citation counts were used to evaluate single scholars.

2. If the cause(s) of an OA advantage were found, it would be useful to expand the research across a broader range of subjects, to see if results were consistent or why variation occurred. Revisiting the subject studied here in several years may provide the opportunity to observe movement in OA levels and potential causes.

3. The study attempted, inconclusively, to see if authors from poorer countries were more likely to cite OA than TA articles. Broadening this question, the citing authors from **1** above should be identified as far as possible to establish whether there is any evidence to suggest that these authors were citing OA articles in preference to TA articles, solely on the free availability of the OA articles. Irrespective of whether this contributed to any citation advantage, such evidence should be used to ensure funder mandates are effective and authors are encouraged as a matter of professional practice to self-archive their work to repositories.

4. Levels of OA varied between subjects. Further research should determine the levels of OA across a range of subjects and regions to identify where there are significant disparities in self-archiving and how these can be remedied.

5. Search tools differed in their recall of OA articles, but the study was limited to four subjects. The coverage of these search tools should be assessed across disciplines and the results used to:

- Rank search tools by general recall and by subject.
- Identify areas of poor coverage.
- Identify coverage by institutional archives.
- Identify the self-archiving practice of authors.
- Assess the usefulness of building a disciplinary archive.

Bibliography

Abel, R. & Newlin, L., eds., 2002. *Scholarly publishing: books, journals, publishers and libraries in the twentieth century*. New York: John Wiley.

About OAIster, [n.d]. <http://OAIster.umdl.umich.edu/o/OAIster/about.html>, [accessed 07.02.07].

About OpenDOAR, 2007. <http://OpenDOAR.org/about.html>, [accessed 31.01.07].

ACS Publications, 2008. <http://pubs.acs.org/rates/institutions/print-2008.pdf>, [accessed 15.04.08].

AGORA, 2005. <http://www.aginternetwork.org/en/>, [accessed 16.2.06].

Anderson, K., Sack, J., Krauss. & O'Keefe, L. 1999. Publishing online-only peer-reviewed biomedical literature: three years of citation, author perception, and usage experience. *The Journal of Electronic Publishing* [online], **6**(3). <http://www.press.umich.edu/jep/06-03/anderson.html>, [accessed 29.09.05].

Allan, G. & Skinner, C., eds. 1991. *Handbook for research students in the social sciences*. London: Falmer Press.

Andrews, J. & Law, D., eds., 2004. *Digital libraries policy: planning and practice.* Aldershot: Ashgate.

Antelman, K., 2004. Do open-access articles have a greater research impact. *College and Research Libraries*, **65**(5), 372-382.

Antelman, K., 2006a. Letter to the Editor: Response to Philip Davis. *College of Research Libraries*, **67**(1), 105.

Antelman, K., 2006b. Self-archiving practice and the influence of publisher policies in the social sciences. *Learned Publishing*. **19**(2), 85-95.

arXiv monthly submission rate statistics, [n.d.]. <http://arxiv.org/show_monthly_submissions>, [accessed 06.03.08]

Ashton, S. & Oppenheim C., 1978. A method of predicting Nobel prizewinners in chemistry. *Social Studies of Science*, **8**(3), 341-348.

Association of Learned and Professional Society Publishers, 2005a. The facts about OA. <http://www.alpsp.org/publications/FAOAcompleteREV.pdf>, [accessed 24.10.05].

Association of Learned and Professional Society Publishers. 2005b. The facts about OA. <http://www.alpsp.org/publications/FAOAaddendum.pdf>, [accessed 17.01.06].

Association of Research Libraries: Create New Systems of Scholarly Communication, 2000. <http://www.arl.org/create/change.html>, [accessed 29.12.05].

Bachrach, S., 2001. SPARC the view from the faculty. *Serials*, (**14**)2, 137.

Baird, L. & Oppenheim, C., 1994. Do citations matter? *Journal of Information Science*, **20**(1), 2-15.

Bakos, Y. & Brynjolfsson, E., 1999. Bundling information goods: pricing profits and efficiency. *Management Science*, 45(**12**), 1613-1630.

Baynes, G., 2005. BioMed Central responds to ALSP's study 'The Facts about OA'. To multiple recipients of list. *American Scientist OA Forum*, 14 October, 16:01:31.

Becker, S. & Bryman, A., 2004. *Understanding research for social policy and practice*. Bristol: Policy Press.

Bence, V. & Oppenheim, C., 2004. Does Bradford-Zipf apply to business and management journals in the 2001 Research Assessment Exercise? *Journal of Information Science,* **30**(5), 469-474.

Bergstrom, T. & Lavaty, R., 2007. *How often do economists self-archive?*. <http://repositories.cdlib.org/ucsbecon/bergstrom/2007a/>, [accessed 30.01.08].

Berlin Declaration on OA to Knowledge in the Sciences and Humanities, [n.d.]. <http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>, [accessed 03.01.06].

Berners-Lee, T., De Roure, D., Harnad. S., Law, D., Murray-Rust P.,Charles Oppenheim, C., Shadbolt, N. & Wilks, Y. 22.08.05. *Journal Publishing and Author Self-Archiving: Peaceful Co-Existence and Fruitful Collaboration*. <http://openaccess.eprints.org/index.php?/archives/20-guid.html>, [accessed 24.08.05].

Bethesda Statement on OA Publishing, [n.d.]. <http://www.earlham.edu/~peters/fos/bethesda.htm>, [accessed 03.01.06].

BioMed Central FAQ, 2006. <http://www.biomedcentral.com/info/about/apcfaq>, [accessed 17.01.06].

BioMed Central OA Charter, 2006. <http://www.biomedcentral.com/info/about/charter>, [accessed 17.01.06].

Blackwell Publishing About Online Open, [n.d.]. <http://www.blackwellpublishing.com/static/onlineopen.asp?site=1>, [accessed 06.03.08].

Bookstein, A., 1976. The bibliometrics distributions. *Library Quarterly*, **46**(4), 416-423.

Borgman, C., ed., 1990. *Scholarly communication and bibliometrics*. London: Sage.

Borgman, C., 2000. *From Gutenburg to the global information structure: access to information in the networked world*. Cambridge, Massachusetts: MIT Press

Borgman, C. & Furner J., 2002. Scholarly communication and bibliometrics. *In:* Cronin, B., ed. *Annual Review of Information Science and Technology*, pp.3-72.

Bosc, H. & Harnad, S., 2005. In a paperless world a new role for academic libraries: providing OA. *Learned Publishing*, **18**(2), 95-99.

Bradford, S., 1948. *Documentation.* London: CrosbyLockwod.

Bradford, S., 1987. *The documentary chaos*. *In:* Meadows, A.J., ed., *The origins of information science,* pp. 113-119.

Brookes, C., 1969. The complete Bradford-Zipf 'bibliograph'. *Journal of Documentation,* **25**(1), 58-60.

Brody, T., Harnad, S. & Carr, L., 2005. *Earlier web usage statistics as predictors of later citation impact.* <http://eprints.ecs.soton.ac.uk/10713/01/timcorr.htm>, [accessed 18.05.05].

Brody, T., Gingras, Y., Hajjem, C., Harnad, S. & Alma Swan. Incentivizing the Open Access Research Web. *CTWatch Quarterly*, **3**(3), [online], <http://www.ctwatch.org/quarterly/articles/2007/08/incentivizing-the-open-access-research-web/>, [accessed 12.11.07].

Bryman A., 2004. *Social research methods*. 2nd ed. Oxford: Oxford University Press.

Burnham, J., 2006. *Scopus* database: a review. *Biomedical Digital Libraries* [online], **3**(1). <http://www.bio-diglib.com/content/3/1/1>, [accessed 24.01.07].

Cameron, J., 2001. Watersheds in scientific journal publishing. *In* Fredrickson, E., ed. *A century of science publishing: a collection of essays.* pp. 245-256.

Carr, L., 2006. Access to self-archive via *Google Scholar* . To multiple recipients of list. *American Scientist OA Forum*, 22 October, 23:56:45 BST.

Carr, L. & Brody, T., 2007. Size isn't everything: sustainable repositories as evidenced by sustainable deposit profiles. *D-Lib Magazine* [online], **13**(7/8). <http://dlib.org/dlib/july07/carr/07carr.html>, [accessed 06.03.08].

Carr, L. & Harnad, S., 2005. *Keystroke economy: a study of the time and effort involved in self-archiving.* <http://eprints.ecs.soton.ac.uk/10688/01/KeystrokeCosting-publicdraft1.pdf>, [accessed 09.01.06]

CERN Document Server, 2008. <http://cdsweb.cern.ch/>, [accessed 07.03.08].

CiteSeer.IST. Scientific Literature Digital Library. <http://citeseer.ist.psu.edu/>, [accessed 08.03.08].

Cox, J., 2003. *Scholarly publishing practice: the ALSP report on academic journal publishers' policies and practice in online publishing*. Worthing: Association of Learned and Professional Society Publishers.

Craig I., Plume, A., McVeigh, M., Pringle, J. & Amin, M. 2007. Do open access articles have greater citation impact? A review of the literature. *Journal of Informetrics*, **1**(3), 239-248.

Creswell, J.W., 2003. *Research design: qualitative, quantitative and mixed methods approaches.* Thousand Oaks Calif: Sage.

Cronin, B., 1984. *The citation process*. Taylor Graham: London.

Cronin, B., 2005. *The hand of science: academic writing and its rewards.* Lanham, Maryland: Scarecrow Press.

Cronin, B. & Barksby Atkins, H., eds., 2000. *The web of knowledge: a festschrift in honor of Eugene Garfield*. Medford, New Jersey: American Society for Information Science and Technology.

Cronin, B., ed. 2002. *Annual Review of Information Science and Technology.* Vol. 36. Information Today: New Jersey.

Cronin, B. & McKenzie, G., 1992. The trajectory of rejection. *Journal of Documentation*, **48**(3), 25-32.

Crow, R., 2002. *The case for institutional repositories: a SPARC position paper*. <http://www.arl.org/sparc/IR/IR_Final_Release_102.pdf>, [accessed 06.01.06].

Davies, E. & Greenwood, H., 2004. Scholarly communication trends – voices from the vortex: a summary of specialist opinion. *Learned Publishing*, **17**(2), 157-167.

Davis, P., 2006. Letter to the Editor: do open-access articles have greater research impact. *College of Research Libraries*, **67**(1), 103-104.

Davis, P., 2007. *PloS Biology – Read Response: Citation advantage of Open Access articles likely explained by quality differential and media effects*. <http://biology.plosjournals.org/perlserv/?request=read-response&doi=10.1371/journal.pbio.0040157#r1438>, [accessed 12.11.07]

Davis, P. & Connolly., 2007. Institutional repositories: evaluating the reasons for the non-use of Cornell univeristy's installation of DSpace. *D-Lib Magazine* [online], **13**(3/4), <http://dlib/march07/davis/03davis.html>

Davis, P. & Fromerth, M., 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, **71**(2), 203-215.

Dblp.uni-trier.d Computer Science Bibliography, 26.01.06. <http://www.informatik.uni-trier.de/~ley/db/welcome.html>, [accessed 26.01.06].

Deis, L. & Goodman, D., (2005). *Web of Science* (2004 version) and *Scopus*. *The Charleston Advisor* [online], **6**(3). <http://www.charlestonco.com/comp.cfm?id=43>, [accessed 23.01.07]

Denscombe. M., 2003. *The good research guide: for small-scale social research projects.* 2nd ed. Maidenhead: Open University Press.

Dess, H., (2006). *Database reviews and reports: Scopus. Issues in Science and Technology Librarianship.* <http://www.istl.org/06-winter/databases4.html>, [accessed 03.05.06].

De Vaus, D., 2001. *Research design in social research.* London: Sage.

De Vries J., 2001. Peer review: the holy cow of science. *In:* Fredrickson, E., ed. *A century of science publishing: a collection of essays*, pp. 231-244.

Diamond, A., [n.d]. *What is a citation worth?* <http://www.garfield.library.upenn.edu/essays/v11p354y1988.pdf>, [accessed 15.2.06].

Dietrich, J., 2007. *The importance of being first: position dependence citation rates on arXiv:astro-ph.* <http://arxiv.org/PS_cache/arxiv/pdf/0712/0712.1037v1.pdf>, [accessed 10.03.08].

Diodato, V., 1994. *Dictionary of bibliometrics*. New York: Haworth Press.

Directory of Open Access Journals, 2006. <http://www.doaj.org/articles/060113>, [accessed 18.01.06].

Directory of Open Access Journals, 2008a. <http://www.doaj.org/articles/about>, [accessed 06.03.08].

Directory of Open Access Journals, 2008b. <http://www.doaj.org/>, [accessed 06.03.08].

Directory of World Repositories, 2008. <http://www.webometrics.info/premierleague.asp>, [accessed 15.02.08].

Drake, M., ed., 2003. *Encyclopedia of Library and Information Science*, 2nd ed. New York: Marcel Dekker.

Duranceau, E., 2004. Electronic journal forum: Cornell and the future of the Big Deal: an interview with Ross Atkinson. *Serials Review*, **30**(2), 127-130.

Easterby-Smith, M. Thorpe, R. & Lowe, A., 2002. *Management research.* London: Sage.

Egghe, L., 2005. Expansion of the field of informetrics: origins and consequences, *Information Processing & Management*, **41**(6), 1311-1316.

Evans, P. & Peters, J., 2005. Analysis of the dispersal of use for journals in Emerald Management Xtra (EMX). *Interlending and Document Supply*, **33**(3), 155-157.

Eysenbach G., 2006. Citation advantage of open access articles. *PLoS Biology*, **4**(5), 692-698.

Eysenbach G., 2007. *PloS Biology – Read Response: Citation advantage of open access articles.* <http://biology.plosjournals.org/perlserv/?request=read-response&doi=10.1371/journal.pbio.0040157#r1438>, [accessed 13.11.07].

Field, A., 2005. *Discovering statistics using SPSS*. 2nd ed. London: Sage.

Frazier, K., 2001. The librarian's dilemma: contemplating the cost of the "Big Deal". *D-Lib Magazine* [online], **7**(3). <http://www.dlib.org/dlib/march01/frazier/03frazier.html>, [accessed 27.12.05].

Fredrickson, E., 2001a. The Dutch publishing scene: Elsevier and North-Holland. *In:* Fredrickson, E., ed. *A century of science publishing: a collection of essay,* pp. 64-68.

Fredrickson, E., ed. 2001. *A century of science publishing: a collection of essays.* Amsterdam: IOS Press.

Friend, F., 2005. The facts about OA journals To multiple recipients of list. *Liblicense.* 14 October, 19:40:44. EDT

Friend, F., 2006. Towards open access to UK research. *In;*Jacobs, N., ed. *Open access: key strategic technical and economic aspects,* pp. 161-167.

Garfield, E., 1979. *Citation indexing: its theory and application in science, technology, and humanities*. New York: John Wiley.

Garfield, E., Of Nobel class: part 2 forecasting Nobel prizes using citation data and the odds against it. *Current Comments*. **35**, 127-136.

Garfield, E., 2005. *The agony and the ecstasy – the history and meaning of the Journal Impact Factor*. <http://garfield.library.upenn.edu/papers/jifchicago2005.pdf>, [accessed 29.01.07].

Garfield, E., [n.d.]. *Can researchers bank on citation analysis?* <http://www.garfield.library.upenn.edu/essays/v11p354y1988.pdf>, [accessed 17.02.06].

Garland, J., 2001. The Dutch publishing scene: Elsevier and North-Holland. *In:* Fredrickson, E., ed. *A century of science publishing: a collection of essay,* pp. 3-14.

Gauch, H., 2003. *Scientific method in practice.* Cambridge University Press: Cambridge.

Gibbs, N., 2005. Walking away from the Big Deal. *Serials*, **18**(20), 89-94.

Ginsparg, Paul., 2003 *Can peer review be better focused?* <http://people.ccmr.cornell.edu/~ginsparg/blurb/pg02pr.html>, [accessed 29.12.05].

Glanzel, W. & Schmoch, U., eds., 2004. *Handbook of quantitative science and technology research*. Dordrecht: Springer.

Goodman, D., Antelman, K. & Bakkalbasi, N., 2005. Identifying OA Articles: Valid and Invalid Methods. *In: Proceedings XXV Annual Charleston Conference: Issues in Book and Serial Acquisition*. <http://dlist.sir.arizona.edu/968/>, [accessed 23.12.05].

Gorman, G. & Rowland. F., eds., 2005. *Scholarly publishing in an electronic era.* London: Facet.

Gower, B., 1997. *Scientific method: an historical and philosophical introduction*. London: Routledge.

Gribbin, J., 2003. *Science a history.* London: Penguin.

Guedon, J., 2001. *In Oldenburg's long shadow: librarians, research scientists, publishers and the control of scientific publishing*. <http://www.arl.org/arl/proceedings/138/guedon.html>, [accessed 01.06.05].

Guedon, J., 2003. OA archives: from scientific plutocracy to the republic of science. *IFLA Journal*, **29**(2), 129-140.

Guedon, J., 2006. Publishing Reform, University Self-Publishing and OA. To multiple recipients of list *American Scientist OA Forum*, 19 Jan, 15:04:45.

Gunnarsdottir, K., 2005. Scientific journal publications: on the role of electronic exchange in the distribution of scientific literature. *Social Studies of Science*, **35**(4), 549-579.

Hair, J., Black, W., Babin, B., Anderson, R. & Tatham R., 2006. *Multivariate data analysis.* 6th ed. New Jersey: Pearson Prentice Hall.

Hajjem, C., 28.07.05. *Étude de la variation de l'impact de citations des articles en accès libre.* <http://www.crsc.uqam.ca/lab/chawki/graphes/EtudeImpact.htm>, [accessed 10.01.06].

Hajjem, C. & Harnad, S., 2007. *Manual evaluation of robot accuracy in automatically identifying open access articles on the web.* <http://eprints.ecs.soton.ac.uk/12220/2/manual-eval.pdf>, [accessed 13.11.07].

Hajjem, C., Harnad, S. & Gringras, Y., [n.d.]. *Ten-year cross-disciplinary comparison of the growth of OA and how it increases citation impact.* <http://eprints.ecs.soton.ac.uk/12906/>, [accessed 10.01.06].

Hajjem, C., Gingras, Y., Brody, T., Carr, L. & Harnad, S. [n.d.]. *Open access to research increases citation impact.* <http://eprints.ecs.soton.ac.uk/11687/>, [accessed 07.12.07].

Harnad, S., [n.d.]. *Maximizing university research impact through self-archiving.* <http://www.ecs.soton.ac.uk/~harnad/Temp/che.htm>, [accessed 05.01.06].

Harnad, S., [n.d.]. *Post-Gutenberg galaxy: the fourth revolution in the means of production of knowledge.* <http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad91.postgutenberg.html>, [accessed 17.02.06].

Harnad, S., 1996. Implementing peer review on the net: scientific quality control in scholarly electronic journals: *In*: Peek, R. & Newby, G., eds. *Scholarly publishing: the electronic frontier,* pp. 103-118.

Harnad, S., 04.07.02. *Scholarly journals at the crossroads: a subversive proposal for electronic publishing*. <http://www.arl.org/scomm/subversive/sub01.html>, [accessed 02.01.06].

Harnad, S., 2003. *Eprints: Electronic Preprints and Postprints* <http://eprints.ecs.soton.ac.uk/7721/1/eprints.htm>, [accessed 12.03.08].

Harnad, S., 2004a. OA to peer reviewed research through author/institution self-archiving: maximising research by maximising online access. *In:* Andrews, J. & Law. D., eds. *Digital libraries policy: planning and practice,* pp. 63-98.

Harnad, S., 2004b. The 1994 "subversive proposal for electronic publishing" at ten. To multiple recipients of list. *American Scientist OA Forum*, 27 June, 14:02:04 BST.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stammerjohanns. H. & Hilf, E. 2004. The access/impact problem and the green and the gold roads to OA. *Serials Review*, **30**(4), 310-314.

Harnad, S., 2005a. *Maximising the Return on UK's Public Investment in Research* <http://eprints.ecs.soton.ac.uk/11220/02/research-rcuk.pdf>, [accessed 21.12.05]
Harnad, S., 2005b. Re: Open access to research worth 1.5bn a year. To multiple recipients of list. *American Scientist OA Forum*, 27 September, 22:22:36 BST.

Harnad, S., 2005c. Re: The Green and Gold Roads to OA. To multiple recipients of list. *American Scientist OA Forum*, 13 December, 13:17:37 GMT.

Harnad, S., 2005d. New international study demonstrates worldwide readiness for OA mandate. To multiple recipients of list *Mailing List SPARC-IR@arl.org Message #321,* 23 June, 19:00:53 GMT.

Harnad, S., 2005e. OA to research worth £1.5bn a year. To multiple recipients of list. *American Scientist OA Forum*, 30 September, 04:30:16 +0100.

Harnad, S., 2005f. The green and gold roads to maximising journal articles access, usage and impact.
<http://www.haworthpress.com/library/StevanHarnad/07012005.asp>, [accessed 17.02 .06].

Harnad, S., 2005g. *OA Impact Advantage = EA + (AA) +(QB) +QA +(CA) +UA*.
<http://eprints.ecs.soton.ac.uk/12085/>, [accessed 06.01.07].

Harnad, S., 2006. Recent manual measures of OA and OAA. To multiple recipients of list. *liblicense-l*, 22 January, 2006 20:41:15 EST.

Harnad, S., 2007a. *The Open Access Citation Advantage: Quality Advantage Or Quality Bias?* <http://openaccess.eprints.org/index.php?/archives/191-The-Open-Access-Citation-Advantage-Quality-Advantage-Or-Quality-Bias.html>, [accessed 13.11.07].

Harnad, S., 2007b. 07.09.07. *Where there's no access problem there's no open access advantage*. <http://openaccess.eprints.org/index.php?/archives/289-Where-Theres-No-Access-Problem-Theres-No-Open-Access-Advantage.html>, [accessed 14.11.07].

Harnad, S., 2008. Harvard adopts 38th green open access self-archiving mandate. To multiple recipients of list. *American Scientist OA Forum*, 13 February, 16:50:16.

Harnad, S. & Brody, T., 2004. Comparing the impact of OA (OA) vs. non-OA articles in the same journals. *D-Lib Magazine* [online], **10**(6), 1-5. <http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html>, [accessed 22.12.05].

Harnad, S. & Hajjem, C., 2007. *Citation advantage for OA self-archiving is independent of journal impact factor, article age and number of co-authors*. <http://eprints.ecs.soton.ac.uk/13329/02/eysen.pdf>, [accessed 05.02.07]

Hart, C., 1998. *Doing a literature review*. Thousand Oaks Calif: Sage.

Heckler, D., 2005. *Occupational employment projections to 2014*. <http://www.bls.gov/opub/mlr/2005/11/art5full.pdf>, [accessed 16.02.06].

Henderson, A., 2002. Diversity and the growth of serious/scholarly/scientific journals. *In:* Abel, R. & Newlin L., eds. *Scholarly publishing: books, journals, publishers and libraries in the twentieth century*, pp. 133-161.

Hertzel, D., 2003. Bibliometrics history. *In*: Drake, M., ed. *Encyclopedia of Library and Information Science*, 2nd ed., pp. 288-328.

Hicks, D. 2004. The four literatures of the social science. *In*: Moed H. Glanzel, W. & Schmoch, U. eds., 2004. *Handbook of quantitative science and technology research*, pp. 473-495.

HighWire Press, [n.d.]. <http://highwire.stanford.edu/about/>, [accessed 07.03.08].

HINARI, 2006. <http://www.who.int/hinari/en/>, [accessed 16.02.06].

Hinton, P., 2004. *Statistics explained*. 2nd ed. London: Routledge.

Hirsch, J., 2005. An index to quantify an individual's scientific research output. *Proceeding of the National Academy of Sciences*. **102**(46), 16569-16572.

Houghton, J., 2005. Economics of publishing and the future of scholarly communication. *In;* Gorman, G. & Rowland. F., eds. *Scholarly publishing in an electronic era*, pp. 165-187.

House of Commons Science and Technology Committee, 2004. *Scientific publications free for all?* London: The Stationery Office.

Huntington, P., Nicholas, D. & Watkinson, A., 2005. Scholarly journal usage: the results of deep log analysis. *Journal of Documentation*, **61**(2), 248-280.

Hutcheson, G. & Sofroniou, N., 1999. *The multivariate social scientist*. London: Sage.
Institutional Archives Registry, 2006**.**
<http://archives.eprints.org/eprints.php?page=all>, [accessed 06.01.06].

Jacobs N., ed., 2006. *Open access : key strategic, technical and economic aspects*. Oxford: Chandos.

Jacso, P., 2004. *Péter's Digital Reference Shelf. Scopus.*
<http://www.galegroup.com/reference/archive/200409/Scopus.html>, [accessed 24.01.07].

Jacso, P., 2005a. *Google Scholar*: the pros and the cons. *Online Information Review*, **29**(2), 208-214.

Jacso, P., 2005b. As we may search – Comparison of major features of *Web of Science*, *Scopus* and *Google Scholar* citation-based and citation-enhanced databases. *Current Science*, **89**(9), 1537-1547.

Jacso, P., 2008. Savvy searching *Google Scholar* revisited. *Online Information Review*, **32**(1), 102-114.

Jeon-Slaughter, H. Hertkovic, A. & Keller, M., 2005. Economics of scientific and medical journals: where do scholars stand in the debate of online journal pricing and site license ownership between libraries and publishers? *First Monday* [online], **10**(3), <http://www.firstmonday.org/issues/issue10_3/jeon/index.html>, [accessed 12.12.05].

Jones, R., Andrew, T & MacColl. J., 2006. *The institutional repository*. Oxford Chandos.

José, B. & Pacios, A., 2005. The impact of consortia purchasing of periodical publications on the document supply service. *Interlending and Document Supply*, **33**(4), 189-195.

Kendal, M., 1960. The bibliography of operational research. *Operational Research Quarterly*, **11**(1/2), 31-36.

Kinnear, P. & Gray, C., 2006. *SPSS 14 made simple.* Hove: Psychology Press.

King, D. & C. Tenopir., 1999. Evolving journals costs: implications for publishers, libraries, and readers. *Learned Publishing*, **12**(4), 251-258.

Kuhn, T., 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kurtz, M., 2004. *Restrictive access policies cut readership of electronic research journal articles by a factor of two*. <http://opcit.eprints.org/feb19oa/kurtz.pdf>, [accessed 22.07.04].

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E.& Murray, S. 2005. The effect of use and access on citations. *Information Processing and Management*, **41**(6), 1395-1402.

Kurtz, M. & Henneken, E., 2007. *Open access does not increase citations for research articles from The Astrophysical Journal.* <http://arxiv.org/ftp/arxiv/papers/0709/0709.0896.pdf>, [accessed 07.09.07].

Kyrillidou, M. & Young, M. 2008. *ARL Statistics 2005-06*. <http://www.arl.org/bm~doc/arlstats06.pdf>, [accessed 14.04.08].

LaGuardia, C., 2005. E-Views and Reviews: *Scopus* vs *Web of Science*. *Library Journal.com*. <http://www.libraryjournal.com/article/CA491154.html%22>, [accessed 23.01.07].

Lamb, C., 2004. OA publishing models: opportunity or threat to scholarly and academic publishers? *Learned Publishing*, **17**(2), 143-150.

Lawrence, S., 2001. Online or invisible. *Nature*, **411**(6837), 521.

Leimkhuler, F., 1967. The Bradford distribution. *Journal of Documentation*, **23**(3), 197-207.

Library and Information Statistics Unit. 2005. *Annual library statistics*. Loughborough: Loughborough University.

Library and Information Statistics Unit. 2006. *Annual library statistics*. Loughborough: Loughborough University.

Liu, M., 1993. Progress in documentation the complexities of citation practice: a review of citation studies. *Journal of Documentation*, **49**(44), 370-408

Lornic, J., 2006. *The bottom line on open access*. <http://www.universityaffairs.ca/issues/2006/march/open_access_01.html>, [accessed 23.02.06].

Lotka, A., 1987. The frequency distribution of scientific productivity. *In:* Meadows, A.J., ed., *The origins of information science,* pp. 113-119.

Lynch, C., 24.09.01. *ARL Bimonthly Report 217 August 2001 Metadata Harvesting and the Open Archives Initiative*. <http://www.arl.org/newsltr/217/mhp.html>, [accessed 08.01.06].

Mabe, M., 2003. The growth and number of journals. *Serials*, **16**(2), 191-197.

Mabe, M. & Amin M., 2001. Growth dynamics of scholarly and scientific journals. *Scientometrics*, **51**(1), 147-162

MacRoberts, M. & MacRoberts, B., 1996. Problems of citation analysis. *Scientometrics*, **36**(3), 435-444.

Marks, R., 2001. Learned societies adapt to new publishing realities. *In:* Fredrickson, E., ed. *A century of science publishing: a collection of essay*, pp. 91-96.

Meadows, A.J., ed., 1987. *The origins of information science.* London: Taylor Graham.

Meadows, A. J., 1998. *Communicating research*. San Diego: Academic Press

Meadows, A. J., 2000. The growth of journal literature: a historical perspective. *In:* Cronin, B. & Barksby Atkins, H., eds. *The web of knowledge*, pp. 87-108.

Metcalfe, T., 2006. The citation impact of digital preprint archives for solar physics papers. *Solar Physics*, 239, 549-553.

Moed, H., 2002. The impact-factors debate: the ISI's uses and limits. *Nature*, **415**(14 February).

Moed, H., 2005. *Citation analysis in research evaluation.* Dordrecht: Springer.

Moed, H., 2007. The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section. *Journal of the American Society for Information Science and Technology*. **58**(13), 2047-2054.

Moed H. K., Glanzel, W. & Schmoch, U. eds., 2004. *Handbook of quantitative science and technology research*. Dordrecht: Springer.

Montgomery, C. & King, D., 2002. Comparing library and user related cost of print and electronic journal collections. *D-Lib Magazine*, [online], **8**(10). <http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/october02/montgomery/10montgomery.html>, [accessed 20.12.05].

Moore, N., 2006. 3rd ed. *How to research*. Facet Publishing: London.

Morris, S., 05.08.05. *ALPSP response to RCUK's proposed position statement on access to research outputs*. <http://www.alpsp.org/2005pdfs/rcuk050805.pdf>, [accessed 22.12.05].

Morris, S., Personal View. 2006. *Learned Publishing*, **19**(1), 73-75.

Myhill, M. (2005). *Google Scholar*. <http://www.charlestonco.com/review.cfm?id=225>, [accessed 23.01.07].

National Institutes for Health: NIH Public Access, [n.d]. <http://publicaccess.nih.gov/policy_development.htm>, [accessed 10.01.06].

Nederhof, A., 2006. Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, **66**(1), 81-10.

Newman, M., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, **46**(5), 323-351.

Nicholas, D., Huntington, P., Russell, B., Watkinson, A., Jamali, H. & Tenopir, C. 2005. The Big Deal: ten years on. *Learned Publishing*, **18**(4), 251-257.

NIH Public Access: September 2005 submission statistics [n.d.]. <http://nihms.nih.gov/stats/2005-09.html>, [accessed 10.01.06].

Norris, M. & Oppenheim, C., 2003. Citation counts and the research assessment exercise V: archaeology and the 2001 RAE. *Journal of Documentation*, **59**(6), 709-730.

Norris, M. & Oppenheim, C., 2007. Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature. *Journal of Informetrics*, **1**(2), 161-169.

Notess, G., 2005. *Scholarly web searching: Google Scholar and Scirus*. <http://www.infotoday.com/Online/jul05/OnTheNet.shtml>, [accessed 23.01.07].

OAIster, 2008. < http://www.oaister.org/>, [accessed 07.03.08].

OAIster statistics. 2007. <http://OAIster.umdl.umich.edu/o/OAIster/stats.html>, [accessed 29.01.07].

Odlyzko, A., 2002. The rapid evolution of scholarly communication. *Learned Publishing*, **15**(1), 7-19.

Okasha, S., 2002. *Philosophy of science*. Oxford: Oxford University Press.

OpenDOAR. 2007. <http://www.OpenDOAR.org/search.php>, [accessed 07.02.07].

Opthof, T., 1997. Sense and nonsense about the impact factor. *Cardiovascular Research*, **33**(1), 1-7.

Oppenheim, C., 1995. The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science University Departments. *Journal of Documentation,* **51**(1), 18-27.

Oppenheim, C**.,** 1997. The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British research in genetics, anatomy and archaeology, *Journal of Documentation,* **53**(5), 477-487.

Oppenheim, C., 2005. OA and the UK Science and Technology Select Committee Report Free for All? *Journal of Librarianship and Information Science*, **37**(1), 3-6.

Oxford Journals, 2006a. <http://www.oxfordjournals.org/oxfordopen/>, [accessed 06.03.08].

Oxford Journals, 2006b. <http://www.oxfordjournals.org/our_journals/nar/announce_openaccess.html>, [accessed 18.02.06].

Oxford Journals Access & Purchase, 2006. <http://www.oxfordjournals.org/access_purchase/self-archiving_policya.html >, [accessed 19.01.06].

Oxford University Press, 2006. <http://www.oup.com/about/>, [accessed 18.02.06].

Page, G., Campbell, R. & Meadows, J., 1997. *Journal publishing.* Cambridge: Cambridge University Press.

Pallant, J., 2005. *SPSS survival manual,* 2nd ed. Maidenhead: Open University.

Peek, R., 1996. Scholarly publishing, facing the new frontiers. *In*: Peek, R. & Newby, G., eds. *Scholarly publishing: the electronic frontier,* pp. 3-15.

Peek, R. & Newby, G., eds., 1996. *Scholarly publishing: the electronic frontier.* Cambridge, Massachusetts: MIT Press

Phillips, E. & Pugh, D., 2001. *How to get a PhD*. 3rd ed. Open University Press: Maidenhead.

Pinfield, S., 2003. Open archives and UK institutions: an overview. *D-Lib Magazine* [online], **9**(3). <http://www.dlib.org/dlib/march03/pinfield/03pinfield.html> [accessed 04.07.05]

Pinfield, S., 2005. A mandate to self-archive? The role of OA institutional repositories. *Serials*, **18**(1), 30-34.

Popper, K., 1968. *The logic of scientific discovery*. Hutchinson: London.

Popper, K., 1992. *Unended Quest*. Routledge: London.

Price, D., 1963. *Little science big science*. Columbia University Press: New York.

Pritchard, A., 1969. Statistical bibliography or bibliometrics? *Journal of Documentation*, **25**(4), 348-349.

Project RoMEO, [n.d.] <http://www.lboro.ac.uk/departments/ls/disresearch/romeo/>, [accessed 11.01.06].
Public Library of Science FAQ, [n.d.]. <http://www.plos.org/faq.html>, [accessed 06.03.08].

Public Library of Science, 2006. <http://www.plos.org/index.html>, [accessed 02.01.06]

Publish Or Perish, 2008. <http://www.harzing.com/resources.htm>, [accessed 02.03.08].

PLoS Journals, [n.d.]. <http://www.plosjournals.org/perlserv/?request=index-html>, [accessed 23.11.07].

PubMed Central Overview, 07.01.06. <http://www.pubmedcentral.nih.gov/about/intro.html>, [accessed 20.01.06].

RAE2008 Research Assessment Exercise, [n.d]. <http://www.rae.ac.uk/pubs/2005/03/>. [accessed 22.12.05].

Reed Elsevier, 2008. <http://www.reed-elsevier.com/index.cfm?articleid=84>, [accessed 06.03.08].

Registry of open access repositories (ROAR), 2008. <http://roar.eprints.org/index.php?action=generate_chart>, [accessed 06.03.08].

Research Councils UK, 2005. RCUK position statement on access to research outputs. <http://www.rcuk.ac.uk/access/statement.pdf>, [accessed 22.12.05].

Research Councils UK., [n.d.]. Access to Research Outputs. <http://www.rcuk.ac.uk/access/default.htm>, [accessed 23.11.07].

UK scholarly journals 2006 baseline report An evidence-based analysis of dataconcerning scholarly journal publishing, 2006. <http://www.rin.ac.uk/files/UK%20Scholarly%20Journals%202006%20Baseline%20Report.pdf>, [accessed 01.02.07].

Research Randomizer, 2007. <http://www.randomizer.org/form.htm>, [accessed 29.01.07].

Reichardt, C. & Rallis, S., eds., 1994. *The qualitative-quantitative debate: new perspectives.* San Francisco: Jossey-Bass.

Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research, 2008. <http://grants.nih.gov/grants/guide/notice-files/not-od-08-033.html>, [accessed 07.03.08].

Richards, S., 1987. *Philosophy & sociology of science: an introduction*. 2nd ed. London: Blackwell.

Robson, C., 2002. *Real world research*. 2nd ed. Oxford: Blackwell.

Roth, D., The emergence of competitors to the Science Citation Index and the *Web of Science*. *Current Science*, **89**(9), 1531-1536.

Rousseau, R., 1994. Similarities between informetrics and econometrics. *Scientometrics*. **30**(2-3), 385-387.

Rowland, F., 2002. The peer-review process. *Learned Publishing*, **15**(4), 247-258.

Rowland, F., Swan, A., Needham, P., Probets, S., Muir, A., Oppenheim, C., O'Brien, A. & Hardy, R. 2004. Delivery, management, and access model for e-prints and OA journals. *Serials Review*, **30**(4), 298-303.

Rowland, F., 2005a. Where is scholarly publishing going? *In*: Gorman, G. & Rowland, F., eds. *Scholarly publishing in an electronic era.* pp. 3-19

Rowland, F., 2005b. Journal access programmes for developing countries. *Serials*, **18**(2), 104-106.

Rowlands, I., 2004. Emerald authorship data, Lotkas's law and research productivity. *Aslib Proceedings: New Information Perspectives*. **57**(1), 5-10.

Rowlands, I., Nicholas, D. & Huntington, P., 2004. Scholarly communication in the digital environment: what do authors want? *Learned Publishing*, **17**(4), 261-273.

Sale, A., 08.12.05. *Comparison of IR content policies in Australia.* <http://eprints.comp.utas.edu.au:81/archive/00000230/01/Comparison_of_content_policies_in_Australia.pdf>, [accessed 11.01.06].

Sarkowski, H., 2001. The growth and decline of German scientific publishing 1850-1945. *In:* Fredrickson, E., ed. *A century of science publishing: a collection of essays,* pp. 25-34.

Schwartz, C., 2005. Reassessing prospects for the OA movement. *College and Research Libraries*, **66**(6), 488-495.

Schwarz, G. & Kennicutt, R., 2004. *Demographic and citation trends in astrophysical papers and preprints.* <http://arxiv.org/PS_cache/astro-ph/pdf/0411/0411275v1.pdf>, [accessed 19 03 07].

Science and Engineering Indicators 2004, 2004. <http://www.nsf.gov/statistics/seind04/c0/c0s1.htm>, [accessed 17.02.06].

Seglen, P., 1992. The skewness of science. *Journal of the American Society for Information Science*, **43**(9), 628-638.

Self-Archiving FAQ, 2005. <http://www.eprints.org/openaccess/self-faq/#self-archiving>, [accessed 16.2.06].

SHERPA Publisher copyright policies & self-archiving: the SHERPA/ROMEO list, [n.d.]. <http://www.sherpa.ac.uk/romeo.php>, [accessed 21.12.05].

SHERPA News. 2006. <http://www.sherpa.ac.uk/news/opendoaroct06.html>, [accessed 31.01.07].

Smith, J., 2007. Re: The apparent OA citation advantage. To multiple recipients of list *JISC Repositories*, 20 May, 19:19:35 BST.

Smith, L., 1981. Citation analysis. *Library Trends*, **30**(1), 83-106.

Smith, A. & Eysenck, M**.,** 2002. *The Correlation between RAE Ratings and Citation Counts in Psychology, Technical Report, Psychology, Royal Holloway College, University of London.* <http://cogprints.org/2749/>, [accessed 28.02.06].

Snyder, H. & Bonzi, S., 1998. Patterns of self-citation across disciplines (1980-1989). *Journal of Information Science,* **24**(6), 431-435.

SPARC about SPARC, 2005. <http://www.arl.org/sparc/about/index.html>, [accessed 20.01.06].

Springer Open Choice, [n.d.]. <http://www.springer.com/sgw/cda/frontpage/0,11855,1-40359-12-115382-0,00.html>, [accessed 06.03.08].

Solomon, D., 2007. *Developing open access electronic journals: a practical guide.* Oxford: Chandos.

Steele, C., 2005. Snap, crackle and ultimately pop? the future of serials. *Serials*, **18**(2), 132-136.

Suber, P., 2005a. *OA Overview.* <http://www.earlham.edu/~peters/fos/overview.htm>, [accessed 02.01.06].

Suber, P., 2005b. *Timeline of OA movement.* <http://www.earlham.edu/~peters/fos/timeline.htm>, [accessed 21.12.05].

Suber, P., 2005c. NIH OA Results Not Encouraging. To multiple recipients of list. *Mailing List SPARC-OAForum@arl.org Message #2472.* 18 October, 08:03:02. <https://mx2.arl.org/Lists/SPARC-OAForum/Message/2472.html>, [accessed 10.01.06].

Suber, P., 2005d. *Open access, impact and demand*. <http://www.bmj.com/cgi/content/full/330/7500/1097>, [accessed 10.12.07].

Suber P., 02.01.06. *OA news.* <http://www.earlham.edu/~peters/fos/fosblog.html>, [accessed 02.01.06].

Summers, R., Oppenheim, C., Meadows, J., McKnight, C. & Kinnell, M. 1999. Information science in 2010: a Loughborough University view. *Journal of the American Society for Information Science* **50**(12), 1153-1162.

Swan, A. & Brown S., 2004a. *JISC/OSI journal authors survey: report*. <http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf>, [accessed 26.10.05].

Swan, A. & Brown S., 2004b. Authors and OA publishing. *Learned Publishing*, **17**(3), 219-224.

Swan, A., 2005. *Open access self archiving: an introduction*. <http://eprints.ecs.soton.ac.uk/11006/>, [accessed 21.02.06].

Swan, A. & Brown S., 2005. *OA self archiving: an author study*. <http://eprints.ecs.soton.ac.uk/10999/01/jisc2.pdf >, [accessed 22.01.07]

Swan, A., Needham, P., Probets, S., Muir, A., Oppenheim, C., O'Brien, A., Hardy, R., Rowland, F. & Brown, S. 2005. Developing a model for e-prints and OA journal content in UK further and higher education. *Learned Publishing*, **18**(1), 25-40.

Tabachnick, B. & Fidell, L., 2001. *Using multivariate statistics.* 4th ed. Boston: Allyn and Bacon.

Tenopir C. & King D., 2000. *Towards electronic journals: realities for scientists, librarians, and publishers*. Washington DC: Special Libraries Association

Testa, J. & McVeigh, M., 2004. *The impact of OA journals: a citation study from Thompson ISI.* <http://scientific.thomson.com/media/presentrep/acropdf/impact-oa-journals.pdf>, [accessed 01.09.05].

The ISI Impact Factor, [n.d.]. <http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/>, [accessed 15.02.06].

The Survey System. 2007. <http://www.surveysystem.com/sscalc.htm#terminology> [accessed 29.01.07].

The Thomson scientific journal selection process, [n.d.]. <http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/>, [accessed 01.03.06].

Thompson, J., 2005. *Books in the digital age*. Cambridge: Polity Press.

Ulrich's Periodicals Directory, 2008. < http://www.ulrichsweb.com/ulrichsweb/>, {accessed 06.03.08].

Van Leeuwen, T. *et al*. 2003. The holy grail of science policy: exploring and combining bibliometrics tools in search of scientific excellence. *Scientometrics*, **57**(2), 257-280.

Van Teijlingen, E. & Hundley, V., 2001. The importance of pilot studies. *Social Research Update* [online], (35), 1-4. <http://www.soc.surrey.ac.uk/sru/SRU35.pdf>, [accessed 26.01.07].

Vaughan, L., 2001. *Statistical methods for the information professional*. American Society for Information Science and Technology: Medford, New Jersey.

Vickery, B., 1948. Bradford's law of scattering. *Journal of Documentation*, **4**(3), 198-203.

Waltham, M., 2005. JISC: *Learned society open access business models.* <http://www.marywaltham.com/JISCReport.pdf>, [accessed 18.02.06].

Waltham, M., 2006. Learned society business models and open access: overview of a recent JISC-funded study. *Learned Publishing*, **19**(10), 15-30.

Ware, M., 2004 *PALS Pathfinder Research on Web-based Repositories: Final report.* <http://aims.ecs.soton.ac.uk/pep.nsf/cc4a508424b9c3ff802566dc004e42ff/5c4d447fc4fdeecf80256e46003c0c0e?OpenDocument>, [accessed 08.01.06].

Web of Science. [n.d.].< http://scientific.thomson.com/products/wos/>, [accessed 01.03.06].

Wellcome Trust., 2003. *Economic analysis of scientific research publishing.* London: Wellcome Trust.

Wellcome Trust., 2005. *Conditions under which a grant is awarded.* <http://www.wellcome.ac.uk/assets/wtx026668.pdf >, [accessed 22.12.05].

Weller, A., 2001. *Editorial peer review: its strengths and weaknesses.* Medford, New Jersey: American Society for Information Science and Technology.

White, H., 2001. Authors as citers over time. *Journal of the American Society for Information Science and Technology*, **52**(2), 87-108.

White, H. & McCain K., 1998. Visualizing a discipline: an author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, **49**(4), 327-55.

Willinsky, J., 2006. *The access principle.* Cambridge, Massachusetts: MIT Press

The World Bank, 2007. <http://go.worldbank.org/K2CKM78CC0>, [accessed 08.08.07].

Wren, J., 2005. Information in Practice. *BMJ* [online], **330**(1128). <http://bmj.bmjjournals.com/cgi/reprint/330/7500/1128>, [accessed 30.06.05].

Zipf, G., 1965. *The psycho-biology of language: an introduction to dynamic philology*. Cambridge, Massachusetts: MIT Press

Appendix A - Pilot studies

Three pilot studies were undertaken to identify whether there was an OA advantage to a sample of articles that were OA, compared to a similar sample of toll access articles. All of the studies were carried out manually; these were:

Study 1 A comparative study similar to that by Antelman (2004, pp.372-382) was carried out in September 2005.

Study 2 A study into a single discipline assessing whether prolific authors were significantly more in evidence than lesser authors in terms of whether or not they had self-archived their work. This work was carried out in November 2005 and January 2006.

Study 3 A comparative study of the citation profile of embargoed journals and entirely closed access journals was carried out in December 2005.

4.2 Study 1

As a preliminary to this study, Kristin Antelman[2] was emailed to obtain guidance and any data which would help this comparative study. Data supplied by her was limited to an Excel spreadsheet which gave brief statistical information and some general comments on web searching. In this study, two subject areas were selected from the *Journal Citation Reports* i.e. sociology and neurology. *Journal Citation Reports* are a product of Thomson ISI they rank a proportion of the journals indexed by the *ISI Web of Knowledge* by their Impact Factor (IF). As indicated earlier the IF of a journal "is calculated by dividing the number of current year citations to the source items published in that journal during the previous two years" (The ISI Impact Factor...[n.d.]). Journals that are ranked in the *Journal Citation Reports* will therefore have had their articles, more or less frequently cited, when compared to other journals in the same subject category. The greater the number of citations relative to the number of articles published in the given period the higher the IF will be. The higher a journal appears in

[2] Kristin Antelman e-mail to Michael Norris, 20 June 2005.

its subject category the greater it is assumed is its influence, authority and prestige. Journals may appear in several subject categories dependent on their coverage.

Six journals titles were taken from the subject category of sociology and three from the subject category neurology. The journal titles are listed below.

Sociology

American Journal of Sociology
British Journal of Sociology
Economy and society
Human Ecology
Law & Society Review
Sociological review

Neurology

Epilepsy Research
Brain
Annals of Neurology

4.2.1 Method

A hard copy of the table of contents for each of the above journals was taken for the year 2003. To ensure only research articles were used in the study, letters to the journal and any reviews were discounted. Using the cited reference search on the *Web of Knowledge* the first author, the journal title and the citing year were entered. This search was done successively for the each authored article, the citations to each article was counted and recorded to an Excel spreadsheet. Any citations which were incorrectly attributed by the giving of an erroneous page or volume number were discarded. The articles title were then copied and pasted as exact phrases into the search engine *Google* in order to find, if possible, a freely available version of the article. If a copy was found, then the article was tagged as having an OA version accessible by anyone. This exercise

was repeated with the beta version of *Google Scholar*. The results from 316 article records were recorded.

4.2.2 Results

The tables below summarise the data collected.

Table 3. Results of Citation Counts for the Neurology Journals

Neurology Journals					
Number articles selected	Articles with no citations	Mean citation rate for cited articles	Range of citations	Mean citation rate for Non OA articles	Mean citation rate for OA articles
102	49	8	1-37	5.5	0.6

Table 4. Results of Citation Counts for the Sociology Journals

Sociology Journals					
Number articles selected	Articles with no citations	Mean citation rate for cited articles	Range of citations	Mean citation rate for Non OA articles	Mean citation rate for OA articles
214	83	2.47	1-12	1.29	1.68

4.2.3 Analysis of results

The results found were not as expected. Antelman found a citation advantage in favour of those authors who had made an OA version of their article available on the web in all the subject areas that were examined. The data collected from those authors publishing their work in the neurological journals without making an OA version available received a greater number of citations, on average, than those who did. There was a citation advantage, albeit small, for those authors in the sociological journals who had published an OA version of their articles to the web. This advantage in percentage terms looks significant at 30% this compares favourably with the work carried out by Antelman where the advantage ranged between 45% and 91%. It compares less well

with the work of Hajjem, Harnad & Gingras (2005) who found a citation advantage of 172% for sociology.

4.3 Study 2

Swan and Brown (2005, p.34) in their latest study which examined the self-archiving habits of authors, found that those authors who were prolific writers in their field were more likely to self-archive than those who were less productive. This assumption, however, was based on the reported behaviour of authors rather than on any analysis of actual author records. This study was intended to test whether Swan and Brown's conclusion was correct, that there is a correlation between author productivity and the number of records self-archived.

4.3.1 Method

A discrete subject, dyslexia, was chosen for examination. A 'general' search within the *Web of Knowledge* revealed there to be 601 articles on this subject (search term dyslex* within article title, articles only) between the years 2000-2006. The results were ranked by author frequency using the 'Analyse' feature within the WoK. From this list, the first 100 records were chosen for analysis. Author productivity by article ranged from 4 to 19. A stratified sample from each quartile of five authors was taken. Only one of the journals was OA. For each author, the number of articles they had published and the number of citations they had received within the period were counted and recorded. The following searches were conducted on each author and article title to determine whether an OA version of the article could be found:

- *OAIster* - search by author and dyslex* in the relevant search fields
- Author's departmental homepage by latest known affiliation – search by author
- *Google* - search using the exact title as a search phrase.

OAIster (2006) is an OAI service provider; it harvests document metadata from institutional and disciplinary archives. Its six million documents include those from, for

example, arXiv, Citebase, Citeseer, Cogprints, CERN, E-LIS and DOAJ as well as a large number of institutional repositories. OA versions of articles that were found at publisher's web sites were ignored as being part of some sort of delayed OA model. Only self-archived versions of the articles were regarded as genuinely OA.

At the first incident of finding an OA version of any one article, the search cycle was curtailed and the next record was selected for processing. Where an OA record was found, if at all, it and its source were recorded. A measure of correlation was carried out between author productivity in terms of the number of articles published and their propensity to make them OA by self-archiving them.

4.3.2 Results

From the sample of 20 authors 149 articles were found of which 30 were OA, some 20% of the total. The 30 OA articles were scattered as shown in Fig 1. There is no obvious correlation between the two sets of values. A correlation calculation showed in fact that there was a very weak negative correlation (-0.06), that is, the fewer the articles published, the more likely that a self-archived version of the article would be found. It was noted that the citation characteristics of several articles did not display the normal decay pattern; they displayed a two hump graphical profile were citations counts had unexpectedly increased.

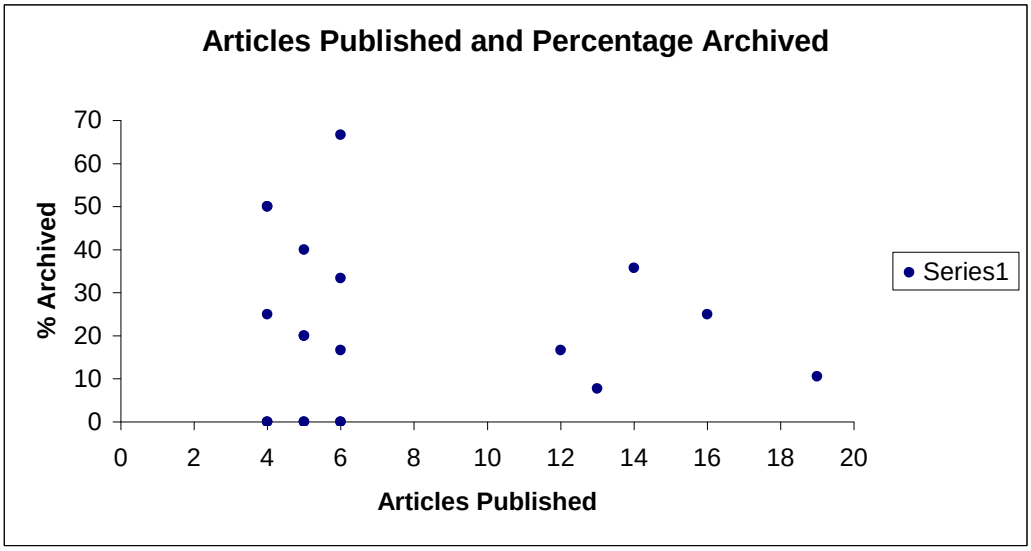


Fig 2. Scatter diagram of Articles Published and % Archived

4.3.3 Analysis

This result does not tie in with the accepted evidence reported by Swan & Brown (2005, p.34). The authors in the survey were of course self-selecting, so may not be representative of the author population as a whole. The negative correlation, although very weak, goes completely against the general speculation that highly productive authors are more likely to self-archive. In fairness, however, Lawrence (2001) and Wren (2005, p.1131) for example do not feel that their research would allow them to say with certainty the cause of the higher citation rates in relation to OA articles and presumably the rates of self-archiving that their research demonstrates. Only Kurtz *et al.* (2005, pp.1398-1401) have felt able to suggest that more highly cited authors were perhaps self-archiving more frequently than those less cited authors.

4.5 Study 3

It had been noted in an earlier pilot study that the citation characteristics of several articles did not display the normal decay pattern; they displayed a two hump graphical profile were citations counts had unexpectedly increased. Examining several of these records showed that the journal from which the articles had been drawn had become OA

after an embargo period had expired. This pilot study addressed the hypothesis that when a toll access journal becomes OA after an embargo period, it will receive more citations than a toll access journal that is entirely closed.

4.5.1 Method

HighWire Press (n.d.) provides access to an aggregated list of online STM journals. Many of these journals allow OA after an embargo period. John Sack[3] of HighWire Press provided a series of historical emails that showed when these journals had first allowed OA and their embargo period. These titles were listed and grouped by embargo period and first OA date. All of the journals were accessed electronically and the OA/TA demarcation tested.

From the *Journal Citation Report* for 2004, a pair of journals were selected from the subject category cell biology, each of comparable impact factor and half life; *Oncology* (IF 6.318, HL 4.6) is toll access and *Journal of Cell Science* (IF 6.91, HL 4.8) is OA after an embargo of period of twelve months. The first notification of the *Journal of Cell Science* becoming OA after an embargo period was in June 2001, with a policy of releasing, in January of each year the previous year's articles, although this later changed in 2004 to becoming OA after 6 months. At January 2002, in this case, all of the journal issues prior to the period from January 2001 would have become OA. The table of contents for each issue of both journals for 1999 was printed and retained. Research articles were identified for analysis; letters and review articles were discarded. The first author, journal title and cited year of each research article was entered into a Cited Reference Search in *WoS*, resulting in a list of citing articles. These were scanned for erroneous citations and then analysed by year and citation count; the results were entered into an Excel spreadsheet. A graph was generated from the data the aggregated citation count was plotted against the year in which the citations were made. To make the two graphs comparable the total citation count for each year was plotted as a percentage of the total citation count for all years.

[3] John Sack e-mail to Michael Norris, 21 October 2005.

4.5.2 Results

The total number of articles in the sample was 849, 436 of which were for the Journal of Cell Science. In all 29621 citations were counted; 16204 were from *Oncogene*. The average number of citations for each Oncogene article was 39 and for the Journal of Cell Science 30. With the exception of one article, all had been cited at least once across the sample period of six years; citation counts ranged between 0 and122 for the *Journal of Cell Science* and 1 and 302 for *Oncogene*. The citation counts for each journal in terms of a percentage of the whole year are given in the table below.

Table 5 Distribution of all citations as a percentage of the total for each year.

	1999	2000	2001	2002	2003	2004	2005
Oncogene	3.7	15.9	20	17.8	15.8	15.4	11.4
Journal Cell Science	2.7	15.4	18.6	17.9	16.9	16.6	12
Difference	1	0.5	1.4	-0.1	-1.1	-1.2	-0.6

A graph generated from the table above is shown below.

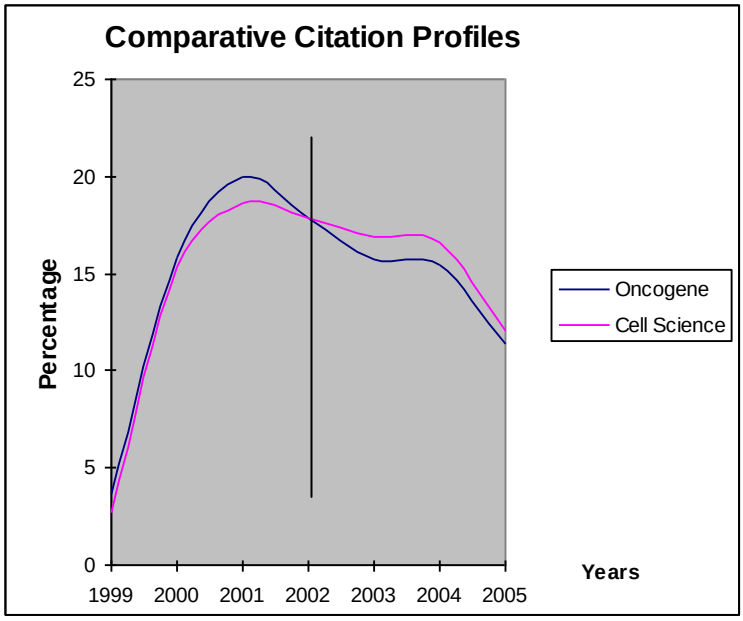


Fig 3 Profile of the two journals by percentage citation counts.

The vertical line on the graph at January 2002 shows the first occasion at which the *Journal of Cell Science* became OA, that is issues prior to January 2001 became freely available.

4.5.3 Analysis

The citation counts varied significantly between journals both in their total and by their average for each journal article. Only one journal article did not receive any citations. This at first seems very unusual, as in the first pilot study described above 42% of the articles did not receive any citations. The period, however, over which these citations counts were accrued was six years and taken from two high impact journals with a similar half-life.

When the yearly citation counts are converted into a percentage for the whole sample period and the two results compared they are very similar in their distribution. The results for the journals show them to be very closely matched, although there is a discernible difference after the embargo period for the *Journal of Cell Science.* Table 5 shows just how small the difference is between the two journals.

In a general study of OA journals Testa & McVeigh (2004) of ISI examined the citation frequency of those OA journals, that are now indexed by them compared to comparable closed access journals. The study did not find a discernible difference in citation count between them. Like this pilot study the ISI study was carried out at the journal level rather than at article level.

4.6 Study Conclusions

From the results found in the three studies undertaken above, it is clear that there is little consistency with the results that have been published to date from similar studies. Results from study 1 showed an OA advantage for sociology but not for neurology. In study 2 it did not appear that the more prolific an author the more likely they were to self-archive; quite the reverse in fact. Study 4 which examined the possibility that an OA advantage would become evident after an embargo period looked as if this might be possible but the advantage is very small and is not sufficiently robust to be thought

significant. One important difference however, between the studies undertaken here and those published, is their scale and statistical significance. By comparison the studies conducted here were small scale, and it is conceded were unsophisticated inasmuch that they were carried out without properly understanding the nature of the scholarly communication process within particular disciplines. This lack of sophistication will need to be addressed. It is proposed to allocate the first two months of the second year to carefully assessing the characteristics of the different disciplines, the nature of the research questions to be asked and the methodology to be used.

The expectation from the pilot studies was that the OA advantage found in the literature would be readily mirrored in the results from these studies. This has proved not to be the case, likewise the pilot study which examined the self-archiving habits of more prolific authors produced an uncharacteristic result when compared to the findings of Swan & Brown (2005, p.34). In year two it is proposed to collect a sufficiently large body of data from which a number of studies can be made, rather than focusing very narrowly on one aspect of OA and limiting data collection to that particular facet. This process will make data collection more extensive and lengthy, but will allow a more flexible use of it. The subject areas from which data will be collected is expected to be different from those studies which have been undertaken already. A number of suggested disciplines which may be examined are given below, subject to the caveat given above on sophistication and methodology:

Economics
Information Science & Library Science
Geology
Microbiology
Optics
Ecology

What is also apparent from the studies undertaken so far is the need for a thoroughly credible and robust methodology that will stand scrutiny. From the work of year one it is evident that the results of Harnad and his colleagues may have been damaged by the criticisms of Goodman and his team, these showed that the results found by Harnad and the computer robot used to retrieve OA records may have been suspect. By extension, it

is not clear if the work by Lawrence, which also used a computer algorithm, is as significant as was first thought. Those studies that appear to be convincing in showing an OA advantage have been smaller in scale and focused on specific disciplines. They have also counted citations manually and as such it is assumed, that the records included have been checked for errors and inconsistencies. Given that manually counting citations in this way lends credibility to the results and that using the *Web of Science* is a well-proven tool for doing this, it is proposed to use this general methodology in the proposed research for year two.

Appendix B - Journal titles

Journal titles and their 2005 impact factor: Applied Maths

Title	Impact Factor	Number of Articles
ACM Transactions on Mathematical Software	1.463	26
Chaos	1.760	123
Communications on Pure and Applied Mathematics	1.841	58
Inverse Problems	1.541	92
Journal of Cryptology	2.280	12
Journal of Mathematical Imaging and Vision	2.197	30
Journal of Non-Linear Science	1.556	18
Journal of Scientific Computing	1.653	41
Mathematical Models and Methods in Applied Sciences	1.248	77
Mathematical Programming	1.497	107
Physica D-Non linear Phenomena	1.863	185
Siam Journal on Applied Dynamical Systems	2.159	23
Siam Journal on Numerical Analysis	1.392	128
Siam Journal on Optimisation	1.238	75

Journal titles and their 2005 impact factor: Applied Maths

Title	Impact Factor	Number of Articles
Siam Journal on Scientific Computing	1.509	143
Siam Review	7.213	20

Journal titles and their 2005 impact factor: Ecology

Title	Impact Factor	Number of Articles
American Naturalist	4.464	138
Conservation Biology	4.110	166
Ecology	4.506	303
Journal of Applied Ecology	4.594	89
Journal of Ecology	4.277	95
Molecular Ecology	4.301	309
Trends in Ecology & Evolution	14.864	71

Journal titles and their 2005 impact factor: Economics

Title	Impact Factor	Number of Articles
Econometrica	2.626	61
Economic Journal	1.440	75
Health Economics	1.919	82
International Economic Review	1.284	54
Journal of Accounting and Economics	1.877	41
Journal of Econometrics	1.579	79
Journal of Economic Geography	3.222	18
Journal of Economic Growth	2.577	13
Journal of Economic Perspectives	2.634	37
Journal of Environmental Economics and Management	1.529	66
Journal of Financial Economics	2.385	60
Journal of Health Economics	2.708	52
Journal of International Economics	1.667	58
Journal of Law and Economics	1.609	24
Journal of Monetary Economics	1.661	64
Journal of Political Economy	2.245	42
Journal of Risk and Uncertainty	2.100	23

Journal titles and their 2005 impact factor: Economics

Title	Impact Factor	Number of Articles
Mathematical Finance	1.345	25
Resource and Energy Economics	1.541	18
Review of Economic Studies	2.035	37
Review of Economics and Statistics	1.518	92
World Development	1.504	120

Journal titles and their 2005 impact factor: Sociology

Title	Impact Factor	Number of Articles
American Journal of Sociology	3.262	58
American Sociological Review	2.933	76
British Journal of Sociology	1.490	62
Economy and Society	1.125	58
Global Networks – A journal of Translational Affairs	1.340	27
Journal for the Scientific Study of Religion	1.039	99
Journal of Marriage and the Family	1.350	159
Language in Society	0.902	40
Law and Society Review	1.396	44
Leisure Sciences	1.045	44
Politics and Society	1.100	36
Population and Development Review	1.076	54
Rural Sociology	1.067	51
Social Networks	1.382	38
Social Problems	1.796	52
Society and Natural Resources	1.339	111

Title	Impact Factor	Number of Articles
Sociological Methods and Research	1.032	31
Sociology of Education	1.222	39
Sociology – The Journal of the British Sociological Association	1.096	85

Appendix C - Outlier details

DiMasi, J., Hansen, R. & Grabowski, H. 2003. The price of innovation: new estimates of drug development costs. *Journal Of Health Economics,* **22**(2), 151-185. **249 citations**

Newman, M., 2003. The structure and function of complex networks. *Siam Review,* **45**(2), 167-256. **639 citations**

Chapman, T., Arnqvist, G., Bangham, J. & Rowe, L Sexual conflict. *Trends In Ecology & Evolution,* **18**(1), 41-47. **168 citations**

Appendix D - Journal titles

Journal titles and their 2005 impact factor: Objective 2 Sociology

Title	Impact Factor	Number of Articles
American Journal of Sociology	3.262	33
American Sociological Review	2.933	37
British Journal of Sociology	1.490	23
Discourse and Society	0.787	32
Economy and Society	1.125	27
Global Networks – A journal of Translational Affairs	1.340	20
Human Ecology	0.909	31
Journal for the Scientific Study of Religion	1.039	32
Journal of Leisure Research	0.791	24
Journal of Marriage and the Family	1.350	82
Language in Society	0.902	40
Law and Society Review	1.396	28
Leisure Sciences	1.045	23
Politics and Society	1.100	21

Journal titles and their 2005 impact factor: Objective 2 Sociology

Title	Impact Factor	Number of Articles
Rural Sociology	1.067	23
Social Forces	1.578	50
Social Indicators Research	0.746	70
Social Networks	1.382	22
Social Problems	1.796	28
Society and Natural Resources	1.339	55
Sociologia Ruralis	1.340	23
Sociological Methods and Research	1.032	15
Sociological Review		29
Sociology – The Journal of the British Sociological Association	1.096	53
Sociology of Education	1.222	14
Sociology of Health and Illness	2.169	42
Theory and Society	0.756	21
Work Employment and Society	1.104	33

Appendix E - Journal titles

Journal titles and their 2005 impact factor: Ecology

Title	Impact Factor	Number of Articles
American Naturalist	4.464	138
Conservation Biology	4.110	166
Ecology	4.506	303
Journal of Applied Ecology	4.594	89
Journal of Ecology	4.277	95
Molecular Ecology	4.301	309

Appendix F - Individual journal citation advantage

First round ecology

Journal title	Citation advantage			
	All citations		Other author Citations	
	OA	TA	OA	TA
American Naturalist	5.41			1.35
Conservation Biology	33.18		36.76	
Ecology	38.37		40.78	
Journal Of Applied Ecology	45.81		32.97	
Journal Of Ecology	41.14		30.40	
Molecular Ecology	36.69		39.79	
Trends In Ecology And Evolution	59.23		57.26	

First round economics

Journal title	Citation advantage			
	All citations		Other author Citations	
	OA	TA	OA	TA
Econometrica	25.34		35.06	
Economic Journal	74.43		72.27	
Health Economics		3.04	0.33	
International Economic Review	215.79		231.58	
Journal Of Accounting And Economics	59.67		72.27	
Journal Of Econometrics	243.92		327.49	
Journal Of Economic Geography	173.15		265.20	
Journal Of Economic Growth	54.29		68.75	
Journal Of Economic Perspectives	92.13		94.51	
Journal Of Environmental Economics And Management	64.84		97.52	
Journal Of Financial Economics	38.46		42.06	
Journal Of Health Economics	7.23		14.17	
Journal Of International Economics	105.43		82	
Journal Of Law And Economics	637.5		543.75	
Journal Of Monetary Economics	136.21		121.38	
Journal Of Political Economy	208.33		207.14	
Journal Of Risk And Uncertainty		8.73		23.61
Mathematical Finance	50		35	
Resource And Energy Economics	220		485.71	
Review Of Economic Studies	102.38		134.38	
Review Of Economics And Statistics		12.32		13.22
World Development	36.26		42.21	

First round applied mathematics

Journal title	Citation advantage			
	All citations		Other author Citations	
	OA	TA	OA	TA
Acm Transactions On Mathematical Software	215.31		247.90	
Chaos	37.87		66.76	
Communications On Pure And Applied Mathematics	4.93			5.13
Inverse Problems	23.24		25.80	
Journal Of Cryptology	466.67		483.33	
Journal Of Mathematical Imaging And Vision	70		41.67	
Journal Of Nonlinear Science	47.06		10	
Journal Of Scientific Computing	211.45		441.67	
Mathematical Models & Methods In Applied Sciences	65.90		124.64	
Mathematical Programming	71.81		59.84	
Physica D-Nonlinear Phenomena	34.89		40.44	
Siam Journal On Applied Dynamical Systems	25.08			13.68
Siam Journal On Numerical Analysis	109.26		201.28	
Siam Journal On Optimization	30.78		10.10	
Siam Journal On Scientific Computing	53.86		79.16	
Siam Review	537.5		983.33	

First round sociology

Journal title	Citation advantage			
	All citations		Other author Citations	
	OA	TA	OA	TA
American Journal Of Sociology	113.56		131.39	
American Sociological Review	74.42		92.64	
British Journal Of Sociology	5.24		6.36	
Economy And Society	21.33		14.04	
Global Networks-A Journal Of Transnational Affairs	140		146.15	
Journal For The Scientific Study Of Religion	61.39		96.43	
Journal Of Marriage And The Family	18.38		12.35	
Language In Society	143.53		194.23	
Law & Society Review	39.07		45.29	
Leisure Sciences	102.5		146.58	
Politics & Society	150.38		187	
Population And Development Review	209.71		200.39	
Rural Sociology		14.53		10.91
Social Networks	20.80		9.90	
Social Problems	94.05		99.88	
Society And Natural Resources	141.48		191.58	
Sociological Methods & Research		30.60		29.72
Sociology Of Education		11.11		1.29
Sociology- The Journal Of The British Sociological Association	46.71		56.24	

Second round economics

Journal title	Citation Advantage			
	All citations		Other author citations	
	OA	TA	OA	TA
Review Of International Political Economy	251.87		228.79	
Quantitative Finance	42.29		11.43	
Post-Soviet Affairs		76.47		90.48
Oxford Economic Papers-New Series		44.27		11.77
Journal Of Urban Economics	43.06		51.75	
Journal Of Economics & Management Strategy	144.67		193.18	
Journal Of Economic Theory	24.45		7.61	
Journal Of Economic Surveys	32.58		23.93	
Journal Of Economic Psychology	43.33		96.55	
Journal Of Economic Behavior & Organisation	70.22		73.92	
Journal Of Development Economics	99.83		124.70	
Journal Of Applied Econometrics	353.33		438.10	
International Journal Of Forecasting	68.62		68.62	
Games & Economic Behavior	96.98		173.82	
Explorations In Economic History	45.83		80.56	
European Economic Review	178.88		242.84	
Economics Of Transition	8.62		22.81	
Econometric Theory	375.89		375.89	
Cambridge Journal Of Economics	74.55		33.33	
Australian Journal Of Agricultural And Resource Economics	184.38		313.64	
American Journal Of Agricultural Economics	56		50.43	

Second round sociology

Journal title	Citation Advantage			
	All citations		Other author citations	
	OA	TA	OA	TA
American Journal Of Sociology	161.40		187.83	
American Sociological Review	75.54		76.92	
British Journal Of Sociology	50		20	
Discourse And Society	46.46			25.64
Economy And Society	107.79		114.29	
Global Networks- A Journal Of Transnational Affairs	296.43		581.82	
Human Ecology	264		146.32	
Journal For The Scientific Study Of Religion	20		13.68	
Journal Of Leisure Research	N/A			
Journal Of Marriage And The Family	41.26		34.26	
Language In Society	233.33		300	
Law And Society Review	86.92		116	
Leisure Sciences	196.88		137.5	
Politics And Society		63.64		33.33
Rural Sociology	50.7		95	
Social Forces	42.22		20.43	
Social Indicators Research	5.93		34.15	
Social Networks	166.67		100	
Social Problems	15.79		12.82	
Society And Natural Resources		22.11		30.17
Sociologia Ruralis	22.33		44.68	
Sociological Methods And Research	43.5		57.5	
Sociological Review		31.36		5.81
Sociology Of Education	15.38		50	
Sociology Of Health & Illness	29.5			10.3

Second round sociology

Journal title	Citation Advantage			
	All citations		Other author citations	
	OA	TA	OA	TA
Sociology- The Journal Of The British Sociological Association		37.06		27.60
Theory And Society	333.33		455.56	
Work Employment And Society	113.33		357.14	

Appendix G - Logistic regression variables

Dependent variable

- Open Access/Not Open Access

Independent variables added to the regression model

- Subject - sociology
- Country of origin by first author affiliation - USA
- Total citations to the article (Totcites)
- Article subject - ecology
- Impact factor of the journal in which the article appeared (IPFactor)
- Country of origin by first author affiliation - Europe
- Number of authors per article (Number of authors)
- Other author citations to the article, excluding self-citations. (OA)

Independent variables discarded by the regression model

- Journal self-citations
- Author self-citations in the same journal
- All author self-citations
- Author self-citations appearing in journals other than the original journal
- Total author and journal self-citations
- Citations from other authors in other journals
- Country of origin by first author affiliation - UK
- Country of origin by first author affiliation - RoW
- Article subject - economics
- Article subject – applied maths

Appendix H - SPSS output

SPSS output for the 'enter' method for variables into the logistic regression model for first round data.

Classification Table[a,b]

Observed			Predicted		
			OA		Percentage Correct
			Toll Access	Open Access	
Step 0	OA	Toll Access	2351	0	100.0
		Open Access	2280	0	.0
	Overall Percentage				50.8

a. Constant is included in the model.

b. The cut value is .500

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients												
			Constant	Numberof authors	IPFactor	Totcites	Tjscites	AASC	ATC	USA	Europe	RoW	Ecology	Economics	Applied_ maths
Step 1	1	5606.150	-1.913	.056	.078	.034	-.046	.159	.014	.535	.192	-.175	.661	1.685	1.554
	2	5577.859	-2.362	.074	.100	.053	-.072	.206	-.002	.663	.267	-.156	.717	1.930	1.809
	3	5577.462	-2.413	.076	.103	.057	-.076	.212	-.006	.675	.275	-.150	.716	1.954	1.835
	4	5577.461	-2.413	.076	.104	.057	-.076	.212	-.006	.675	.276	-.150	.716	1.954	1.835

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 6418.841

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

e. Redundancies in Design Matrix:
OA = Tjscites - AASC
OAC = Totcites - Tjscites + AASC - ATC
Sociology = 1 - Ecology - Economics - Applied_maths

AOC = -AASC + AT
Tot_Self_cites
uk = 1 - U

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	841.379	12	.000
	Block	841.379	12	.000
	Model	841.379	12	.000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.797	8	.559

Classification Table[a]

			Predicted		
			OA		Percentage
	Observed		Toll Access	Open Access	Correct
Step 1	OA	Toll Access	1446	905	61.5
		Open Access	614	1666	73.1
	Overall Percentage				67.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Numberofauthors	.076	.025	9.259	1	.002	1.079	1.027	1.132
	IPFactor	.104	.027	14.202	1	.000	1.109	1.051	1.170
	Totcites	.057	.007	67.092	1	.000	1.058	1.044	1.073
	Tjscites	-.076	.030	6.272	1	.012	.927	.873	.984
	AASC	.212	.081	6.830	1	.009	1.236	1.055	1.450
	ATC	-.006	.022	.073	1	.787	.994	.952	1.038
	USA	.675	.110	37.905	1	.000	1.963	1.584	2.434
	Europe	.276	.124	4.958	1	.026	1.317	1.034	1.679
	RoW	-.150	.140	1.156	1	.282	.861	.655	1.131
	Ecology	.716	.140	26.022	1	.000	2.045	1.554	2.693
	Economics	1.954	.099	391.441	1	.000	7.057	5.815	8.564
	Applied_maths	1.835	.104	311.177	1	.000	6.266	5.110	7.683
	Constant	-2.413	.144	280.197	1	.000	.090		

a. Variable(s) entered on step 1: Numberofauthors, IPFactor, Totcites, Tjscites, AASC, ATC, USA, Europe, RoW, Ecology, Economics, Applied_maths.

SPSS output for the 'backward stepwise' method for variables into the logistic regression model for first round data.

Classification Table[a,b]

Observed			Predicted		
			OA		Percentage Correct
			Toll Access	Open Access	
Step 0	OA	Toll Access	2351	0	100.0
		Open Access	2280	0	.0
	Overall Percentage				50.8

a. Constant is included in the model.

b. The cut value is .500

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients										
			Constant	Numberof authors	IPFactor	Totcites	Tjscites	AASC	USA	Europe	Ecology	Economics	Applied_ maths
Step 3	1	5607.610	-1.996	.058	.078	.036	-.049	.182	.618	.282	.662	1.682	1.544
	2	5579.055	-2.443	.074	.101	.053	-.070	.202	.744	.352	.704	1.929	1.793
	3	5578.709	-2.491	.075	.105	.056	-.073	.202	.754	.358	.701	1.952	1.817
	4	5578.709	-2.491	.075	.106	.056	-.073	.202	.754	.358	.700	1.952	1.818

a. Method: Backward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 6418.841

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

e. Redundancies in Design Matrix:
OA = Tjscites - AASC
OAC = Totcites - Tjscites + AASC - ATC
Tjscites - AASC + ATC
Sociology = 1 - Ecology - Economics - Applied_maths

RoW = 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	841.379	12	.000
	Block	841.379	12	.000
	Model	841.379	12	.000
Step 2[a]	Step	-.073	1	.787
	Block	841.306	11	.000
	Model	841.306	11	.000
Step 3[a]	Step	-1.174	1	.279
	Block	840.132	10	.000
	Model	840.132	10	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.797	8	.559
2	6.557	8	.585
3	8.405	8	.395

Classification Table[a]

			Predicted		
			OA		Percentage
	Observed		Toll Access	Open Access	Correct
Step 1	OA	Toll Access	1446	905	61.5
		Open Access	614	1666	73.1
	Overall Percentage				67.2
Step 2	OA	Toll Access	1447	904	61.5
		Open Access	613	1667	73.1
	Overall Percentage				67.2
Step 3	OA	Toll Access	1453	898	61.8
		Open Access	630	1650	72.4
	Overall Percentage				67.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 3[a]	Numberofauthors	.075	.025	9.331	1	.002	1.078	1.027	1.131
	IPFactor	.106	.027	14.839	1	.000	1.111	1.053	1.173
	Totcites	.056	.006	91.774	1	.000	1.057	1.045	1.070
	Tjscites	-.073	.029	6.181	1	.013	.930	.878	.985
	AASC	.202	.073	7.694	1	.006	1.224	1.061	1.413
	USA	.754	.083	83.172	1	.000	2.126	1.808	2.500
	Europe	.358	.098	13.390	1	.000	1.430	1.181	1.733
	Ecology	.700	.139	25.430	1	.000	2.015	1.534	2.645
	Economics	1.952	.099	390.822	1	.000	7.045	5.805	8.549
	Applied_maths	1.818	.102	315.842	1	.000	6.157	5.038	7.523
	Constant	-2.491	.125	396.986	1	.000	.083		

a. Variable(s) entered on step 1: Numberofauthors, IPFactor, Totcites, Tjscites, AASC, ATC, uk, USA, Europe, Ecology, Economics, Applied_maths.

Appendix I – Publications

Norris, M., Oppenheim, C. & Rowland, F., (2008 In press). The citation advantage of open access articles. *Journal of the American Society for Information and Technology*.

The Citation Advantage of Open Access Articles

Michael Norris, Charles Oppenheim and Fytton Rowland
Department of Information Science, Loughborough University, Loughborough, LE11 3TU, UK. E-mail: {M.Norris2, C.Oppenheim, J.F.Rowland}@lboro.ac.uk
Tel +44 (0) 1509 223065 Fax +44 (0) 1509 223053
Corresponding author Charles Oppenheim

Abstract
Four subjects, ecology, applied mathematics, sociology and economics, were selected to assess whether there is a citation advantage between journal articles that have an open access (OA) version on the Internet compared to those articles that are exclusively toll access (TA). Citations were counted using the Web of Science and the OA status of articles was determined by searching OAIster, OpenDOAR, Google and Google Scholar. Of a sample of 4633 articles examined, 2280 (49%) were OA and had a mean citation count of 9.04, whereas the mean for TA articles was 5.76. There appears to be a clear citation advantage for those articles that are OA as opposed to those that are TA. This advantage, however, varies between disciplines, with sociology having the highest citation advantage but the lowest number of OA articles from the sample taken and ecology having the highest individual citation count for OA articles but the smallest citation advantage. Tests of correlation or association between OA status and a number of variables were generally found to be weak or inconsistent. The cause of this citation advantage has not been determined.

Introduction
Academics are frequently judged, in part at least, on the quality of their published research. The greater the impact of that research as counted by, for example, the number of citations it receives, the better, it is believed, is the quality of the work (van Leeuwen

et al. 2003, pp. 262-263). Receiving many citations for academic research generally correlates strongly with academic success; an analysis of Nobel laureates and their citation counts by Garfield (1979, pp. 63-64) and Opthof (1997, p. 2), although tenuous, gives some credibility to the idea that the two are linked. Likewise a similar ranking by Hirsch (2007, pp. 16569-16572) using his *h-index*, which uses article citation counts, has been successfully used to identify and rank prominent physicists.

In recent years, it has become possible for authors to self-archive an electronic version of their work in a variety of locations from personal web pages, to a disciplinary archive or to an institutional repository. In so doing, authors make their work open access (OA) and freely available to anyone who has Internet access. Toll access (TA) articles (often known as closed access articles) remain behind subscription barriers and are only accessible by a personal or institutional subscription. If it can be shown that self-archived OA research often receives more citations than closed access research, then a convincing argument can be made to persuade researchers to self-archive their work. A number of studies (Antelman, 2004; Eysenbach, 2006; Harnad & Brody, 2004; Lawrence, 2001) have shown that those authors who make their work OA will receive more citations and hence achieve greater impact than those authors whose articles remain behind subscription barriers. Despite this advantage, few authors, however, self-archive their work (Swan & Brown, 2005, pp. 62-68); the self-archiving rate of authors seems to be at best, around 15% (Hajjem et al. 2005; Sale, 2005). However, increasing citation impact is not the only benefit of self-archiving; work that is freely available for anyone to read should increase research access and impact, and allow those that fund research through their taxation access as well (Harnad, 2006, p. 73).

There has been much discussion on the causes of this citation advantage. Kurtz et al. (2005), Davis and Fromerth (2006) and Moed (2006) are not convinced that simply making an article open access is sufficient cause for any increased citation counts. Rather, they suggest that authors may self-archive their better quality work, and because some articles are made available as preprints before publication, they have a longer period in which to attract citations. Metcalfe (2006, p. 549), however, thinks that, in solar physics at least, higher citation rates are not the result of authors archiving their higher quality papers, or necessarily that better authors more readily archive their work. What is evident, however, despite what might be the cause of any OA citation

advantage, is that the evidence accumulated so far indicates that those authors who make their peer-reviewed work more visible by self-archiving their articles receive more citations than those who do not.

Previous research

Open Access Citation Advantage

Lawrence (2001) was the first to show that conference articles that were OA and freely available on the World Wide Web were more frequently cited than articles that were offline. Since Lawrence's pioneering work, there have been a number of studies that have demonstrated a similar citation advantage (Antelman, 2004; Eysenbach, 2006; Hajjem et al. 2005; Harnad, 2004). Harnad and his colleagues (Hajjem et al. 2005; Harnad & Brody, 2004; Harnad et al. 2004) have carried out large-scale trials where they examined the citation counts of OA and TA articles from the same journals from a database of 14 million articles. In physics and in a range of other subjects, they have found a significant citation advantage for those articles that were OA. In these studies, they identified OA versions of articles either by trawling the web using a computer algorithm or by taking self-archived versions from a disciplinary archive and then compared the citation counts of both OA and TA versions. In contrast to this approach, Antelman (2004) selected four subjects and a relatively small number of articles and manually identified OA versions of articles and their respective citation counts. Again, there was a significant citation advantage for those articles that were OA, but with noticeable variations between subjects.

These two approaches counted the citations from work that was made available by authors by self-archiving their work where it could be accessed. Eysenbach (2006) took a selection of articles that appeared in a single journal (*Proceedings of the National Academy of Sciences*), some of which were TA and others which were OA by virtue of their authors paying for their publication, even though after a six months moratorium all articles appearing in the journal become OA. The OA articles were available from the publisher's web site. Overall, Eysenbach found, that even when taking into account factors such as the number of authors, country of origin and discipline, that OA articles were still twice as likely to be cited as the non-OA articles appearing in the same

journal. Eysenbach (2006, p. 697) also suggested, that those OA articles that were hosted on the publisher's website were more heavily cited than some of the original TA articles which were subsequently made OA by being self-archived by their authors elsewhere. Given the status of the *Proceedings of the National Academy of Sciences* as a prestigious journal with a high rejection rate and high impact, the results found by Eysenbach are not necessarily applicable to journals containing articles of a more variable quality.

Causation

Lawrence (2001) found a significant correlation between conference papers that were available online and their greater citation counts as compared to offline papers, he was unable to identify the cause of this correlation, although he did suggest in his analysis of papers from the same conferences that "online articles are more highly cited because of their easier availability" (p.521). This uncertainty as to the cause of any OA citation advantage has led to speculation that there are other reasons for this advantage other than simply that the article is OA. Several possible reasons have been suggested, including article age, number of authors, the quality of articles, and the status of authors or of their institutions. Kurtz et al. (2005) looked at three possible reasons for this advantage in astronomy. They found evidence for an early access (EA) effect caused by an article preprint being made freely available prior to journal publication, and a self-selection bias (SB) where the author has self-archived their better work; but were unable to find a specific open access (OA) effect. They concluded, in astronomy at least, that this lack of an OA effect was probably caused because authors in astronomy must already have access to the literature in order for them to carry out and report their research. Wren (2005) found that articles from high-impact biomedical journals are more likely to be found at non-journal websites, suggesting, possibly, that these are better quality papers which are made more readily available by their authors.

Although working with a small sample and a single journal, Eysenbach (2006, p. 697) thought that "…publishing papers as OA articles on journal sites facilitates knowledge dissemination to a greater degree than self-archiving…". This view that self-archiving is less efficient in terms of accruing citations is contentious. Harnad (2007a) has reported the preliminary findings in which his team have quantified four components of a

citation advantage from biomedical articles that had been self-archived. It was shown "that each of the four factors contributes an independent, statistically significant increment to the citation counts" (Harnad, 2007b). The largest increment to any OA citation advantage was the number of years since publication, followed by the impact factor of the journal in which article appeared, the number of authors of the article and that, although the smallest contributor, the fact that the article had been self-archived. Davis and Fromerth (2006), taking article-level data from four mathematics journals, 18.5% of which had been deposited in the arXiv archive, could only find reasonable evidence to support a quality differential where more highly citable articles had been deposited in arXiv. Using a similar approach, Moed (2006) looked to estimate the early view and quality bias effect on the citation impact of preprint articles found in the condensed matter section of arXiv. Taking a large sample from 24 journals of deposited and non-deposited articles, Moed found a strong early view and quality bias, but was unable to find a general open access citation advantage. In a recent review of the literature Craig et al. (2007) could find no evidence of an OA effect, rather they suggested that an article's OA "status alone had little or no effect on citations". The authors supported the work of Moed (2006) which they regarded as the most rigorous to date and if replicated, they argued, this might help determine the generality of the results found by Moed.

Metcalfe (2006) compared the citation rates of solar physics articles made freely available in arXiv or in the Montana State University archive found a citation advantage compared to those articles that had not been deposited. More interestingly, Metcalfe (2006, p. 551) suggested that this effect is due to improved visibility rather than authors selecting their better papers to archive. Metcalfe noted the results of Schwarz and Kennicutt (2004) who found that astrophysics conference papers posted to arXiv were cited twice as frequently as those that were not. Metcalfe sampled a set of conference proceedings from solar physics and found a comparable boost in citation rates for those that had been self-archived to arXiv. Metcalfe (2006, p. 551) suggests that conferences in astronomy and astrophysics are not affected by a quality bias because they are the place to publish work in progress or details that are not significant enough by themselves to merit publication in a peer-reviewed journal, and so he concludes are of lesser quality.

Research background

The present study extends the range of subjects examined by Antelman (2004) and on a smaller scale supplements the work of Hajjem et al. (2005). Four subjects are examined to see if there is an OA citation advantage from articles published in a range of high impact journals. Subject differences are investigated in their level of OA and citation advantage, and the sources of the citations are broken down into, for example journal self-citations and author self-citations to examine any effect on OA advantage. Some measure of causation of the OA effect is made by examining correlations between the number of authors and their articles, by examining the country of origin of OA authors and their particular subjects.

Methods

Harnad and Brody (2004) argued that the best way to test for a citation advantage for OA articles is to compare the citation counts of individual OA and non-OA articles appearing in the same non-OA journal. This process is dependent on these articles accruing citations which can be counted and compared. Given that as many as 50% (Garfield, 2005) of articles are not cited at all, choosing high impact factor (IF) journals from which to take a sample of articles should increase the likelihood that there is a substantial citation count for both TA and OA articles. The impact factor of a journal is calculated by taking the number citations to all documents published in a journal over a consecutive two year period and then dividing this count by the number of citable items from that journal during the same period (Garfield 1979). This metric is calculated annually by Thomson ISI which then ranks the journals it indexes on this basis within subject categories. This data is made available by Thomson ISI in its *Journal Citation Reports* (JCR) and the bibliographic data and accruing citation counts associated with the articles within the indexed journals appears in its citations indexes found in its Web of Science (WoS) database. Moed (2005, pp. 113-114) describes the advantages of using the WoS citation indexes, not least of which is the frequency with which they have been used by other researchers. Coverage of subjects within WoS varies between disciplines, with the sciences predominating. The database is, however, sufficiently broad to enable records to be collected from a range of subjects. Disciplines vary in their level of citedness and the coverage of the subject by journals as opposed to

coverage by monographs; sociology is a particular example where monographs play a significant part in scholarly communication (Nederhof, 2006, pp. 83-86).

Four subjects were selected for examination; these were: applied mathematics; ecology; economics; and sociology. The subjects represent a selection from the sciences and the social sciences. Moed (2005, pp. 126-131) ranks the ISI coverage of these subjects by the number of references made to articles published in a sample of up to eight ISI source journals relative to the total number of references which appear in those source journals. Ecology has the highest coverage at 64%, followed by applied mathematics at 54%, economics at 47% and sociology at 27%. Sociology emerges typically, as Hicks (2004, pp. 480-484) describes, as a discipline which is biased towards publishing a significant amount of material in monographs, leading to a lower number of citations to core journals; that is, authors cite significantly fewer journal articles than, say, in ecology. Both sociology and economics have a relatively high national orientation, as defined by the share of the papers from the country most frequently publishing in a journal, relative to the total number of papers published in the journal (Moed, 2005, p. 131). This suggests that articles published in these subjects are of interest locally to the country rather than being of international significance. Generally speaking, the 'harder' the science, the more likely that scholarly communication will be through journals that are more international in scope. For example, chemistry has an ISI coverage of 84% and a low national orientation (Moed, 2005, pp. 129-131).

A deliberately purposive sampling approach was adopted in the selection of journal titles, since the aim of the work was to assess whether there was an OA citation advantage, and not to determine whether the distribution of OA articles was random or otherwise. By their very nature, high impact journals attract a greater number of citations than their lower impact counterparts and are more likely to have articles from leading academics and their institutions. Whilst the sample was clearly biased in favor of high impact journals, the citation counts attracted by articles from them, would reflect the citing behavior of that particular discipline and hence allow comparisons between them, and a measure of any OA citation advantage if present. A sample of high impact journals from the four disciplines as defined by subject categories in the 2005 edition of the JCR was taken. Appendix A gives the journal titles, their subject category and impact factor. Checking the publisher's websites of the 65 chosen journals showed

that they were all available electronically and with the exception of three, they were also available in print format. The status of each journal was checked to ensure that it was completely TA and that it was not available in OA form after any embargo period.

The bibliographic details and citation counts of all the articles published in 2003 for the journals selected were taken from the citation indexes on WoS. In the case of sociology, because of the high number of book reviews in the journals, a small number of article records and their citation counts were taken from the latter part of 2002 to increase the sample size. This approach was adopted to give rough parity in terms of journal impact factor between the subjects selected without having to take article records from mid-range impact journals and to give a similar sample size to the other subjects. Letters, editorial material and corrections were excluded. In this process, 4633 articles were identified. Moed (2005, p. 95) demonstrates that, in general, the peak in citation frequency for journal articles is usually achieved by about the third year after publication, but there is some variation in this between disciplines. The citation counts from these records were broken down into journal and author self-citations and other author citations. Finding OA versions of TA articles on the World Wide Web can sometimes be difficult and misleading. Searches using computer algorithms, although manageable, can give some false drops (Hajjem et al. 2005). Similarly, manual searches using search engines can lead unwittingly to publisher's websites, especially if searches are made from subscribing institutions where the IP address is recognised by the publisher's server. Additionally, given the many spurious hits that a search engine can give, finding OA versions of articles may prove difficult, even if in fact they are there.

The majority of earlier studies looking for an OA citation advantage have searched the web either manually or by trawling using a computer algorithm. Carr (2006) reports that, of those making searches on the WWW for articles, over 96% of these searchers get to the Eprints repository at Southampton University by using Google (76.05%), Google Scholar (15.25%) and Yahoo (4.93%). The use of Google Scholar to find OA articles has not as yet, we believe, been reported in the literature. In a number of papers, Jacso (2005a, pp. 208-214; 2005b, pp. 1537-1547) has reviewed Google Scholar, from which it is clear that it is not currently an adequate tool for citation counting as such, but it is useful in locating OA versions of journal articles. The view that Google Scholar has limitations is also shared, generally, by others who have also found significant

omissions in the coverage and recall from this database (Myhill, 2005; Notess, 2005). In recent research conducted by Meho and Yang (2007) they found that Google Scholar was particularly strong in its coverage of conference proceedings and international non-English language journals, despite its evident limitations. Whilst these criticisms of Google Scholar are fair, a recent test of its recall by Norris and Oppenheim (2007) found that in terms of finding links to individual articles taken from a common database of articles from the social sciences, Google Scholar had a hit rate of 87%, compared to 88% for WoS and 95% for Scopus.

Additionally, in a small pilot study, a hundred article records were taken from different subjects and used as a sample in the search engines Yahoo, Google and Google Scholar. Each article title was entered as a phrase in each of the search engines. Yahoo was not as successful at finding hits as was Google or Google Scholar, nor did Yahoo find any hits in addition to those found by Google or Google Scholar. However, Google and Google Scholar had little overlap and did return unique hits that found OA article records, suggesting that in combination they would yield more OA results than if used singly.

There has been a significant growth in the number of institutional repositories into which authors can self-archive their journal output and make it freely available, thus broadening the availability of OA material. These repositories can have their records harvested by service providers such as OAIster. OAIster is a union catalogue of digital sources hosted at the University of Michigan (OAIster 2007). Repositories make their records available to OAIster, who then harvest their descriptive metadata using OAI-PMH (the Open Archives Initiative Protocol for Metadata Harvesting). OAIster currently harvests records from over 700 repositories and contains over 10 million records, which are searchable from a single access point. OpenDOAR, (OpenDOAR 2007) hosted by the University of Nottingham, is a similar centralised access point to worldwide institutional repositories. OpenDOAR, initially a directory of open access repositories, now offers a trial service to search the contents of the repositories that it lists. Unlike OAIster, OpenDOAR does not search the repositories' metadata even if they are OAI-PMH compliant, but relies on Google's indexes, and repositories being suitably structured for the Googlebot web crawler. Both of these service providers enable access to well known repositories, including the major subject repositories such as arXiv and RePEc. RePEc (http://repec.org/docs/RePEcIntro.html) is "the world's

largest collection of scholarly information for economics" and its database holds details of 12,700 professionals and 10,250 institutions associated with economics". For our research, Google, Google Scholar, OAIster and the OpenDOAR service were used in combination as the search tools to maximise the findings of OA versions of journal articles.

To determine the OA status of the 4633 articles identified, article titles were entered as a phrase search in OAIster, OpenDOAR, Google Scholar and Google. The search sequence started with OAIster,and proceeded through OpenDOAR, and if necessary Google Scholar, and finally Google. OAIster and OpenDOAR were always searched; if these two failed to produce a hit, then Google Scholar was searched, and finally, if necessary, Google was used. OA articles were those articles that could be identified as being completely freely available from an individual's website, a departmental site, subject repository or an institutional repository. Such finds included preprints, postprints and drafts, and were counted as OA if both the title of the article and the article's authors were the same as that found in the journal in which the article was published. Hits that led to a publisher's website were discounted, as in general a subscription is needed to access the full text.

Findings

Of the 4633 articles tested, 2280 were OA and in total, the 4633 articles accrued 34,156 citations between them; 489 of the articles did not receive any citations at all. Of the 489 articles that did not receive any citations 309 (63.2%) were TA and the remaining 180 (36.8%) articles were OA. Overall, including zero citation count records, the gross mean citation count for those articles that were OA was 9.04 compared to 5.76 for the TA articles. This represents an OA citation advantage of 57% (OA-TA/TA citation counts). When journal and author self-citations were excluded, the mean citation counts for the two article sets were OA 6.47 and TA 3.93, an OA citation advantage of 64%. Figure 1 shows the proportion of OA/TA articles found by subject.

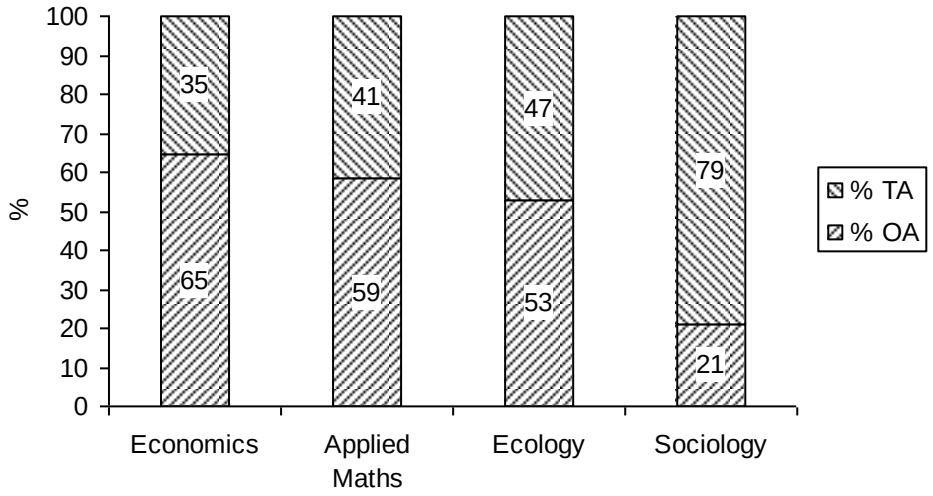


FIG. 1. Proportion of OA/TA articles by subject

The mean citation counts of the two populations, both gross and net of journal and author self-citation counts, were compared using the independent 2 sample *t*-test; the result showed them to be from populations with different means ($p<0.001$). Similarly when the test was conducted for each of the four subjects, the same result was found. Although the frequency distributions of citation counts are usually skewed, such distributions can, when the sample size is sufficiently large, be considered to have means normally distributed in accordance with the central limit theorem (Hinton, 2004, p. 55). The non-parametric Kolmogorov-Smirnov Z test was also used to confirm that the two groups, OA and TA articles, were not drawn from the same populations, and in every instance, the test confirmed that this was the case.

Table 1 gives the gross citation counts for the four subjects; the OA advantage ranges from 88% for sociology to 44% for ecology.

Table 1. Gross citation counts

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	1627	480	3.39	3518	678	5.19	53
Ecology	6240	553	11.28	10012	618	16.20	44
Economics	1716	402	4.27	5099	739	6.90	62
Sociology	3961	918	4.31	1983	245	8.09	88
Total	13544	2353	5.76	20612	2280	9.04	57

This advantage is maintained when journal and author self-citations are removed, leaving just the citations from other authors writing in journals other than the cited article journals; this is shown in Table 2.

Table 2. Citation count net of author and journal self-citations

	TA citations	TA articles	Avg citations TA article	OA citation	OA articles	Avg citations OA article	OA % advantage ±
Applied maths	854	480	1.78	2065	678	3.05	71
Ecology	4246	553	7.68	7058	618	11.42	49
Economics	1245	402	3.10	4056	739	5.49	77
Sociology	2891	918	3.15	1568	245	6.40	103
Total	9236	2353	3.93	14747	2280	6.47	65

Sociology has the highest citation advantage for those articles that are OA, but overall as shown in Figure 1 its authors make the smallest number of their articles OA. Ecology, with the second lowest rate of open access, has the highest citation count for its articles. Economics has the highest rate of OA adoption and is second to sociology in citation advantage.

Figure 2 shows a breakdown of the OA citation count by the four types identified. Journal author self-citations (JASC) are where the cited author is citing themselves and writing in the same journal as the original cited article. Journal self-citations (JSC) are citations where authors other than the original article author have cited the article within the same journal. Author self-citations (ASC) are where the authors are citing themselves but are writing in a journal other than the journal in which their original article appeared. Finally, other citations (OC) are from authors unrelated to the original cited journal or any of its authors.

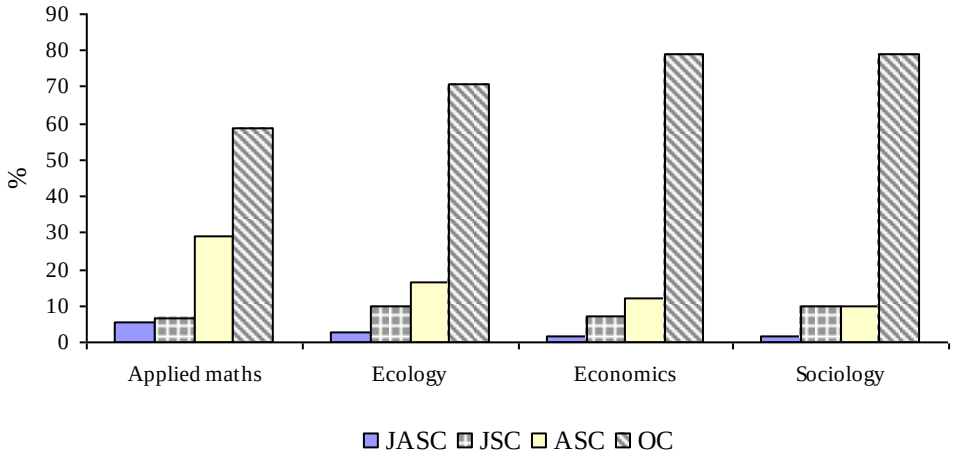


FIG. 2. Breakdown of OA citations by subject

Other author citations form the largest single category for all the subjects and author self-citation rate is highest in applied mathematics and lowest in sociology. The combined self-citation rates are 41% for maths, 29% ecology, and 20% for both economics and sociology. The mean number of journal and author self-citations for OA articles was 2.57, and for TA articles, this was 1.83. Consistent differences between the mean number of journal and author self-citations were also evident at subject level and are consistently greater for OA articles than TA articles; the mean OA/TA journal and author self-citation counts were respectively for ecology 4.78/3.61; economics 1.41/1.17; applied mathematics; 2.14/1.61; and sociology 1.69/1.17. These means were compared using the independent 2 sample t-test; the result showed all four to be from populations with different means ($p<0.001$).

The mean number of authors for all of the articles was 2.34. For all TA articles the mean number of authors was 2.21 and for OA articles this was 2.46. At subject level in every case OA articles had a slightly higher mean number of authors than TA articles. Table 3 gives a breakdown of the mean author counts by subject.

TABLE 3. OA/TA counts by country and subject.

		Subject and article count				
		Ecology	Economics	Applied Math	Sociology	Total
Open Access	N America	383	520	293	190	1386
	Europe	121	95	254	21	491
	UK	65	69	43	24	201
	Rest of World	49	55	88	10	202
Total		618	739	678	245	2280
Toll Access	N America	236	221	159	621	1237
	Europe	147	65	189	65	466
	UK	80	72	23	149	324
	Rest of World	90	44	109	83	326
Total		553	402	480	918	2353

The results of a Chi-square test ($\chi^2(8) = 88.83$, $p < .001$) showed there was a significant association overall between the number of authors and the OA/TA status of an article. However, the association between the number of authors and the OA status of an article showed that there was a tendency towards OA status only when there was more than one author. Hence, there is a strong association between single authorship and articles being TA. Of the 1356 single authored articles 61.36% were TA and the remaining 38.64% were OA. The situation is reversed for articles having more than one author, with these articles having a tendency towards OA status. For those OA articles having two to five authors, the differences between them and the TA articles is, however, less marked and ranges, dependent on the number of authors, from 53%-56% to the advantage of OA articles. This result however, breaks down at subject level where the association is not significant for ecology or sociology, but is for applied mathematics and economics ($p < 0.001$), although even this association is not entirely consistent throughout the range of author counts for the TA and OA articles.

Examining the origin of articles by first author affiliation shows that authors from North America provided the majority of articles, accounting for 57% of all the articles published in the 65 journal titles. Figure 3 gives a breakdown of the OA status of articles by country and subject. North America has the highest rate of OA (60.8%) followed by continental Europe (21.5%), with the UK and the Rest of the World at 8.8% and 8.9% respectively.

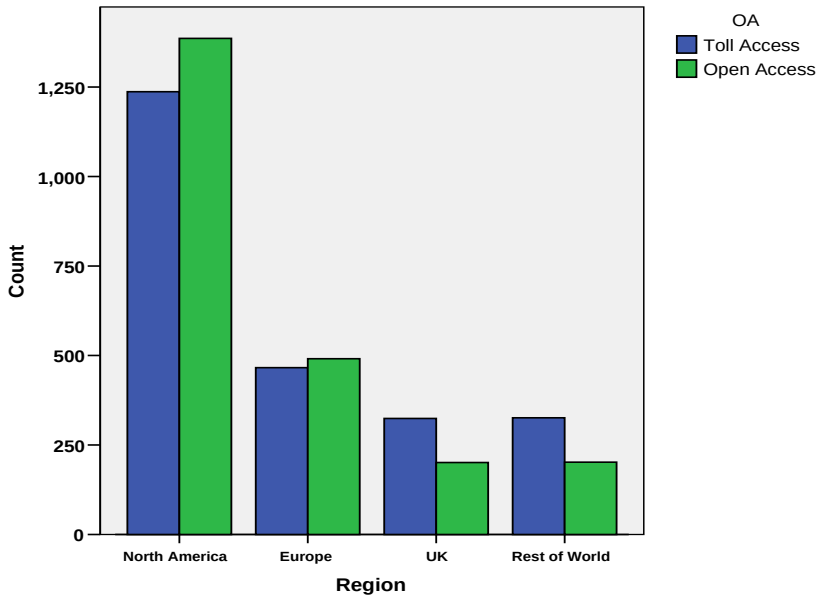


FIG 3. Number of OA/TA articles by region

The results of a Chi-square test ($\chi^2(3) =$, $p<0.002$) showed there was a significant association overall by region, subject and split between OA and TA articles. There is a tendency towards OA in North America and continental Europe, and a tendency toward TA in the UK and the Rest of the World.

The correlation between the number of authors and the total number of citations was tested. For OA articles, this was 0.19 and for TA articles, this was 0.21. Whilst these correlations were significant ($p<0.01$) given the relatively large sample size, the actual results were poorly correlated. Taking just journal and author self–citations and comparing this total to the level of authorship revealed no substantial differences between the two sets of data; the correlations were 0.309 and 0.288 for OA and TA articles respectively ($p<0.01$). Similarly, correlation between journal impact factor and the number of authors was for OA 0.16 and for TA 0.25.

Search engine success

As a by-product of the data gathering, the success of OAIster and OpenDOAR in retrieving OA versions of articles was measured. In comparison to the success of Google and Google Scholar, their success overall were relatively poor. Only in economics and applied mathematics could OAIster and OpenDOAR be considered a

relative success, finding 21% and 22% respectively of the hits between them. A breakdown of hits shows that, of the total 2280 OA items, only 14% overall were found between OAIster and OpenDOAR, with Google and Google Scholar finding the other 86%. A particularly useful feature of Google Scholar was the way that it grouped multiple finds of an article into a single hit and from it, if present, an OA article could readily be found without having to search several pages of records.

A difficulty faced by all web searches is the consistency of web links to, in this case, OA articles. Many articles which look to have viable OA web addresses have broken links, and hence were counted on the day of interrogation as TA even though on another search they may appear as OA. In the case of Google, its results are relatively consistent. In a study of web citations by Vaughan and Shaw (2005, p. 1078), the stability of their initial search results from Google were checked by subsequent repeat searches and found to be fairly constant. The other side of this problem is failing to find OA versions of articles when they are, in fact, available. This appears at first as a positive feature because not finding an OA article would suggest that the OA citation advantage is being understated. However, it can be argued that an OA article that is hard to find and remains unfound will have a lower citation count than those that are easily found, and by default will become coded as TA; if this effect applied in large numbers of cases, it would artificially widen the citation advantage of OA articles. This is an issue for all the research that has been undertaken so far, and it is argued here that searches for OA articles through two general search engines, the metadata of an international repository and a surrogate search of international repositories through Google's indexes have minimised this problem.

Discussion and Conclusion

Out of the 2280 OA articles identified, 86% were found using Google or Google Scholar; at 14%, the finds by OAIster and OpenDOAR were relatively modest, suggesting that the majority of open access articles are not deposited in institutional archives where either OAIster or OpenDOAR can find them. For these subjects at least, in these higher impact journals, the best strategy to find an OA article would be to use Google Scholar followed by Google and then use OAIster or OpenDOAR.

The results found in this work agree with earlier studies that have examined the broad citation advantage of a range of subjects where OA articles are dispersed across the Internet rather than confined to a single subject repository. Notably, the work reported by Antelman (2004), Harnad et al. (2004) and Hajjem et al. (2005) has given similar results. Other work which has found an OA citation advantage has either concentrated on the results from a single journal (Eysenbach, 2006) or particular subjects which have a preprint culture, such as high energy physics, and almost exclusively, in the case of arXiv, their own repository.

The results show a statistically significant difference in the mean number of citations that OA articles received when compared to TA articles. This is apparent for all of the subjects for both the gross citation count and when journal and author self-citations are removed. There are however, variations in both the degree of OA and the citation advantage within the four subjects, with sociology having the smallest number of articles that are OA but having the highest citation advantage. Similarly, Hajjem et al. (2005) reported from their large-scale study that sociology had, at 172%, the largest citation advantage from the ten subjects they examined, although unlike the result here, it also had the highest OA rate as well. In the results here, economics had, at 65%, the highest OA rate and this, it is suggested, is related to the frequency with article metadata is deposited in RePEc and the frequency with which this is found through OAIster and OpenDOAR and hence the ease with which the work can be accessed. Antelman (2004) reported an OA frequency for mathematics of 69% compared to a similar result for applied mathematics of 59% in the
results here.

The majority of citations that the articles received were not author or journal self-citations, although in the case of mathematics a substantial number of them were (41%). In all subjects, however, OA articles were self-cited more frequently than TA articles. This however, does not account for the overall gross OA mean citation advantage over TA articles. Indeed the OA citation advantage is even more marked when self-citations are removed from both sets of counts.

Overall, there is a significant association between the number of authors an article has and whether it is OA or not; generally single authored articles are more likely to be TA.

The results, however, become inconclusive when considered at subject level, and for example, there is no significant association between author count and OA/TA status in sociology.

Most articles originated from North America (57%) when first author affiliation was used to identify their origin. Although a little mixed, there was a noticeable bias in favor of authors making there work OA in North America and continental Europe, this was evident at subject level as well. This was not the case for the UK or the Rest of the World where OA rates were generally lower; however, applied mathematics had almost consistently more OA articles than TA articles in all four regions, and the reverse was true for sociology. Despite this relatively poor position for sociology, its OA citation advantage was the highest, suggesting that where scholars can find what few articles are OA, they are cited heavily. In a similar finding to Antelman (2005, p. 377), who found in her sample that mathematics had the highest number of OA articles and that it also had the lowest citation advantage, our results show that applied mathematics had the second highest number of OA articles, but the third lowest citation advantage.

Other measures of association or correlation were generally inconclusive, leaving the issues of causation of any OA citation advantage unclear. Whilst there was an obvious association between single authors and TA status, this was much less decisive when there was more than one author. Likewise, measures of correlation between impact factor and OA status were found to be weak as was the correlation between the number of authors an article had, and the number of citations it received.
It is evident that the level of OA is subject dependent, and that within these subjects, there are different levels of authorship and citation practices thereby making it difficult to explain the cause of any OA citation advantage. The idea that early access to preprint articles is an explanation for OA citation advantage is not proved, since unlike articles posted to arXiv, the subjects we examined are less well served by a recognised subject repository, except possibly RePEc for economics. Likewise, solely ascribing the advantage to a quality bias is difficult to sustain, since with the exception of sociology, well over half of the articles were OA. As Harnad (2007b) suggests, it is likely to be combination of factors.

Although the reasons why there is a citation advantage for OA articles has still not been satisfactorily explained, it is clear that the advantage exists and occurs regularly across a range of subject areas. Further data collection is planned to investigate the possible cause of this advantage. This may allow some conclusions to be drawn on the reasons for any OA advantage.

References

Antelman, K. (2004). Do open-access articles have a greater research impact?. College and Research Libraries, 65, 372-382. Retrieved March 22, 2007, from http://eprints.rclis.org/archive/00002309/

Carr, L. (2006, October 22). Access to self-archive via Google Scholar. Message posted to http://listserver.sigmaxi.org/sc/wa.exe?A2=ind06&L=american-scientist-open-access-forum&D=1&O=D&F=l&S=&P=81868

Craig, I., Plume, A., McVeigh, M., Pringle, J., & Amin M. Do open access articles have greater citation impact? A critical review of the literature.

Retrieved July 21, 2007, from http://www.publishingresearch.net/Citations-SummaryPaper3_000.pdf.pdf

Davis, P., & Fromerth, M. (2006). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Retrieved March 22, 2007, from http://arxiv.org/abs/cs.DL/0603056

Eysenbach G. (2006). Citation advantage of open access articles. PLoS Biology, 4, 692-698. Retrieved March 25, 2007, from http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040157&ct=1

Garfield, E. (1979). Citation indexing – its theory and application in science, technology, and humanities. New York: John Wiley.

Garfield, E. (2005). The agony and the ecstasy – the history and meaning of the Journal Impact Factor. Retrieved January 8, 2007, from http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

Hajjem, C., Harnad, S., & Gringras, Y. (2005). Ten-year cross disciplinary comparison of the growth of OA and how it increases citation impact. Retrieved January 8, 2007, from http://eprints.ecs.soton.ac.uk/11688/01/hajjem.pdf

Harnad, S. (2006). Opening access by overcoming Zeno's paralysis. In N. Jacobs (Ed.), Open access: Key strategic, technical and economic aspects (pp. 73-85). Oxford: Chandos Publishing. . Retrieved April 3, 2007, from http://eprints.ecs.soton.ac.uk/11688/01/hajjem.pdf

Harnad, S. (2007a). The open access citation advantage: quality advantage or quality bias?. Retrieved April 3, 2007, from http://openaccess.eprints.org/index.php?/archives/191-The-Open-Access-Citation-Advantage-Quality-Advantage-Or-Quality-Bias.html

Harnad, S. (2007b). Citation advantage for OA self-archiving is independent of journal impact factor, article age, and number of co-authors. Retrieved April 5, 2007, from http://openaccess.eprints.org/index.php?/archives/2007/01/17.html

Harnad, S., & Brody, T. (2004). Comparing the impact of OA (OA) vs. non-OA articles in the same journals. D-Lib Magazine, 10. Retrieved March 23, 2007, from http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., & Hilfet, E. (2004). The access/impact problem and the green

and gold roads to open access. Serials Review, 30, 310-314. Retrieved April 5, 2007, from http://users.ecs.soton.ac.uk/harnad/Temp/impact.html

Hicks, D. (2004). The four literatures of the social science. In H. K. Moed., W. Glanzel, & U. Schmoch (Eds.), Handbook of quantitative science and technology research (pp. 473–495). Dordrecht: Springer.

Hinton, P. (2004). Statistics explained. (2nd ed.). London: Routledge.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 46, 16569-16572. Retrieved April 5, 2007, from http://www.pnas.org/cgi/reprint/102/46/16569

Jacso, P. (2005a). Google Scholar: the pros and the cons. Online Information Review, 29, 208-214.

Jacso, P. (2005b). As we may search – Comparison of major features of Web of Science, *Scopus* and *Google Scholar* citation-based and citation-enhanced databases. Current Science, 89, 1537-1547. Retrieved April 5, 2007, from http://www.pnas.org/cgi/reprint/102/46/16569

Journal Citation Reports (2007). Retrieved January 8, 2007, from http://portal.isiknowledge.com/portal.cgi?DestApp=JCR&Func=Frame

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. (2005). The effect of use and access on citations**.** Information Processing and Management, 41, 1395–1402

Lawrence, S. (2001). Online or invisible. Nature, 411, 521. Retrieved April 11, 2007, from http://citeseer.ist.psu.edu/lawrence01online.html

Metcalfe, T. (2006). The citation impact of digital preprint archives for solar physics papers. Solar Physics, 239, 549-553. Retrieved April 12, 2007, from http://www.springerlink.com/content/3485x525622j0801/fulltext.pdf

Meho, L., & Yang K. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. Journal of the American Society for Information Science and Technology. [In press] . Retrieved July 22, 2007, from http://www.slis.indiana.edu/faculty/meho/meho-yang-03.pdf

Moed, H. (2005). Citation analysis in research evaluation. Dordrecht: Springer.

Moed, H. (2006). The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section. Retrieved March 20, 2007, from http://arxiv.org/abs/cs.DL/0611060

Myhill, M. (2005). *Google Scholar*. Retrieved February 15, 2007, from http://www.charlestonco.com/review.cfm?id=225

Nederhof, A. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. Scientometrics, 66, 81-100.

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature. Journal of Informetrics, 1, 161-169.

Notess, G. (2005). Scholarly web searching: Google Scholar and Scirus. Retrieved February 15, 2007, from http://www.infotoday.com/Online/jul05/OnTheNet.shtml

OAIster. (2007). Retrieved January 29, 2007, from http://OAIster.umdl.umich.edu/o/OAIster/about.html

OpenDOAR (2007). Retrieved January 29, 2007, from http://www.opendoar.org/about.html

Opthof, T. (1997). Sense and nonsense about the impact factor. Cardiovascular Research, 33, 1-7.

The RePEc Project (2007). Retrieved July 22, 2007, from http://repec.org/docs/RePEcIntro.html

Sale, A. (2005) Comparison of IR content policies in Australia. Retrieved February 12, 2007, from http://eprints.comp.utas.edu.au:81/archive/00000230/01/Comparison_of_content_policies_in_Australia.pdf

Schwarz, G., & Kennicutt, R. (2004). Demographic and citation trends in astrophysical papers and preprints. Retrieved March 19, 2007, from http://arxiv.org/PS_cache/astro-ph/pdf/0411/0411275v1.pdf

Swan, A., & Brown S. (2005). OA self archiving: an author study. Retrieved March 26, 2007, from http://eprints.ecs.soton.ac.uk/10999/01/jisc2.pdf

Van Leeuwen, T. N., Visser, M. S., Moed, H.F., Nederhof, T.J., & Van Raan A. F. J. (2003). The holy grail of science policy: exploring and combining bibliometrics tools in search of scientific excellence. Scientometrics, 57, 257-280.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: a comparison of four disciplines. Journal of the American Society for Information Science and Technology, 56, 1075-1087.

Web of Science. (2007). Retrieved March 23, 2007, from http://portal.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame

Wren, J. 2005. Information in Practice. BMJ, 330. Retrieved March 23, 2007, from http://bmj.bmjjournals.com/cgi/reprint/330/7500/1128

Appendix A

Journal titles and their 2005 impact factors: Applied Mathematics

Title	Impact Factor
ACM Transactions on Mathematical Software	1.463
Chaos	1.760
Communications on Pure and Applied Mathematics	1.841
Inverses Problems	1.541
Journal of Cryptology	2.280
Journal of Mathematical Imaging and Vision	2.197
Journal of Non-Linear Science	1.556
Journal of Scientific Computing	1.653
Mathematical Models and Methods in Applied Sciences	1.248
Mathematical Programming	1.497
Physica D-Non linear Phenomena	1.863
Siam Journal on Applied Dynamical Systems	2.159
Siam Journal on Numerical Analysis	1.392
Siam Journal on Optimisation	1.238
Siam Journal on Scientific Computing	1.509
Siam Review	7.213

Journal titles and their 2005 impact factors: Ecology

Title	Impact Factor
American Naturalist	4.464
Conservation Biology	4.110
Ecology	4.506
Journal of Applied Ecology	4.594
Journal of Ecology	4.277
Molecular Ecology	4.301
Trends in Ecology & Evolution	14.864

Journal titles and their 2005 impact factors: Economics

Title	Impact Factor
Econometrica	2.626
Economic Journal	1.440
Health Economics	1.919
International Economic Review	1.284
Journal of Accounting and Economics	1.877
Journal of Econometrics	1.579
Journal of Economic Geography	3.222
Journal of Economic Growth	2.577
Journal of Economic Perspectives	2.634
Journal of Environmental Economics and Management	1.529
Journal of Financial Economics	2.385
Journal of Health Economics	2.708
Journal of International Economics	1.667
Journal of Law and Economics	1.609
Journal of Monetary Economics	1.661
Journal of Political Economy	2.245
Journal of Risk and Uncertainty	2.100
Mathematical Finance	1.345
Resource and Energy Economics	1.541
Review of Economic Studies	2.035
Review of Economics and Statistics	1.518
World Development	1.504

Journal titles and their 2005 impact factors: Sociology

Title	Impact Factor
American Journal of Sociology	3.262
American Sociological Review	2.933
British Journal of Sociology	1.49
Economy and Society	1.125
Global Networks – A journal of Translational Affairs	1.340
Journal for the Scientific Study of Religion	1.039
Journal of Marriage and the Family	1.350
Language in Society	0.902
Law and Society Review	1.396
Leisure Sciences	1.045
Politics and Society	1.100
Population and Development Review	1.076
Rural Sociology	1.067
Social Networks	1.382
Social Problems	1.796
Society and Natural Resources	1.339
Sociological Methods and Research	1.032
Sociology of Education	1.222
Sociology – The Journal of the British Sociological Association	1.096

Open Access Citation Rates and Developing Countries

Michael Norris[1]; Charles Oppenheim[1]; Fytton Rowland[2]

[1]Department of Information Science, Loughborough University,
Loughborough, Leicestershire LE11 3TU, UK
e-mail: M.Norris2@lboro.ac.uk; C.Oppenheim@lboro.ac.uk
[2]73 Dudley Street, Bedford MK40 3TA, UK
e-mail: fytton@googlemail.com

The ELPUB 2008 Conference

Abstract

Academics, having written their peer reviewed articles, may at some stage in the make their work Open Access (OA). They can do this by self-archiving an electronic version of their article to a personal or departmental web page or to an institutional or subject repository, such that the article then becomes freely available to anyone with Internet access to read and cite. Those authors who do not wish to do this may leave their article solely in the hands of a toll access (TA) journal publisher who charges for access, consigning their article to remain behind a subscription barrier. Lawrence (2003), in a short study, noted that conference articles in computer science that were freely available on the World Wide Web were more highly cited that those that were not. Following this, there have been a number of studies which have tried to establish whether peer-reviewed articles from a range of disciplines which are freely available on the World Wide Web, and hence are OA, accrue more citations than those articles which remain behind subscription barriers (Antelman 2004, Davis and Fromerth 2007, Eysenbach 2006, Harnad and Brody 2004, Kurtz and Henneken 2007, Moed 2007). These authors generally agree that there is a citation advantage to those articles that have been made OA, but are either uncertain about, or find that they cannot agree on, the cause of this advantage. The causes of this citation advantage could simply be that OA articles are available well in advance of formal publication, and so have a longer period in which to accrue citations, or simply that more authors, because they are freely available, can read and cite them. As part of this debate, Smith (2007) asked whether authors from developing countries might contribute to higher citation counts by accessing OA articles and citing them more readily than TA articles. As part of a larger study of the citation advantage of OA articles (Norris, Oppenheim and Rowland 2008), research was undertaken to see whether a higher proportion of citations to OA articles came from authors based in countries where funds for the purchase of journals are very limited. Mathematics was chosen as the field to be studied, because no special programme for access in developing countries, such as HINARI (2007), covers this subject. The results show that the majority of citations were given by Americans to Americans, but the admittedly small number of citations from authors in developing countries do seem to show a higher proportion of citations given to OA articles than is the case for citations from developed countries. Some of the evidence for this conclusion is, however, mixed, with some of the data pointing toward a more complex picture of citation behaviour.

Keywords: Open Access, Citation advantage, Developing countries

1. Introduction

One of the basic arguments for OA is that those who cannot afford access to peer-reviewed journal articles could access them if the authors of these articles self-archived their work somewhere on the World Wide Web. It should follow that a higher percentage of those who cite these OA articles ought to come from countries where access to expensive journals is limited. A number of schemes, such a HINARI (2007) and AGORA (2007), exist to provide access to scholarly information inexpensively to users in developing countries, but not all disciplines are covered by these. In the overall larger study (Norris, Oppenheim and Rowland 2008), four subjects (sociology, economics, ecology and mathematics) were selected, and a large number of papers were investigated to discover whether they were available on an OA basis anywhere. Citation data on all these papers were collected and subjected to statistical analyses of various kinds to establish whether or not OA availability of itself correlates with a greater number of citations to an article. As part of the larger study, mathematics – which is not covered by any

of the assistance schemes – was chosen for an investigation of citation of articles by authors based in developing countries, the hypothesis being that these authors would be unlikely to have access to expensive toll access (TA) journals.

2. Methods

In the main project, articles were selected from high-impact journals and their OA status was sought by using various search tools (OAIster 2007, OpenDOAR 2007, Google Scholar, and finally Google), and their availability or non-availability with OA was noted. Citations to them were then retrieved by using the ISI Web of Science databases.

The country of origin of cited and citing articles was decided by the first author's affiliation. Countries were classified by their per capita income using the World Bank's (2007) categories (see Table 32), and were also grouped by their geographical location into twelve groups (see Table 50). Citation ratios for the TA group and the OA group of articles were calculated separately.

3. Data

In the overall sample, 1158 mathematics journal articles were taken from 16 high impact journals. Only citation links from other-author citations were counted; all other types of author and journal self-citations were discarded. After this, 365 of the articles were then uncited, leaving 793 articles cited by other authors. Table 50 shows how these 793 were distributed amongst the twelve regions.

Table 50. Cited articles by region and OA status

			OA		
			Toll Access	Open Access	Total
	Spain	Count	12	14	26
		% within Region	46.2%	53.8%	100.0%
	Japan	Count	16	9	25
		% within Region	64.0%	36.0%	100.0%
	Italy	Count	20	22	42
		% within Region	47.6%	52.4%	100.0%
	Germany	Count	18	47	65
		% within Region	27.7%	72.3%	100.0%
	France	Count	26	39	65
		% within Region	40.0%	60.0%	100.0%
	Canada	Count	11	16	27
		% within Region	40.7%	59.3%	100.0%
	Pacific Rim	Count	9	22	31
		% within Region	29.0%	71.0%	100.0%
	China	Count	15	11	26
		% within Region	57.7%	42.3%	100.0%
	Rest of World	Count	20	18	38
		% within Region	52.6%	47.4%	100.0%
	UK	Count	18	28	46
		% within Region	39.1%	60.9%	100.0%
	Europe	Count	31	65	96
		% within Region	32.3%	67.7%	100.0%
	USA	Count	102	204	306
		% within Region	33.3%	66.7%	100.0%
Total		Count	298	495	793
		% within Region	37.6%	62.4%	100.0%

All of the citation links to the remaining 793 articles, which totalled 3032, were then analysed. These 3032 citations were from 2680 citing articles; clearly, in some cases, there were multiple citations from some of the citing articles. Table 51 shows how the 3032 citations from the 2680 citing articles that cited the original 793 were broken down. For example, there were 2413 citing articles (80%) that cited just one of the 793 articles, whereas there were three articles which cited six of the original articles each. The first-author affiliations of the original 793 cited articles covered 47 countries. The first-author affiliation of the 2680 citing articles, citing the 793 articles, were drawn from 70 countries; 23 of these were necessarily in addition to the initial 47 countries. The cited and citing countries were classified by their per capita income using The World Bank's (2007) system of classification. China, for example, is designated as being in the lower middle-income group of countries and India in the low-income bracket, whereas most of Western Europe and North America are in the high-income group of countries. To further aid analysis and comparison, the original 47 countries and the 70 citing-author countries were classified by location into USA, Canada, France, Germany, Italy, Japan, Spain, UK, rest of Continental Europe, China, Pacific Rim, and the Rest of World.

Table 51. Frequency of Citation

Frequency of Citation	Citing Articles	Overall Citations
1	2413	2413
2	208	416
3	43	129
4	9	36
5	4	20
6	3	18
Totals	2680	3032

In

the 3032 citations are shown by their cited article OA or TA status, the region from which they were cited, and whether the cited article was matched by a citation from the same region. By way of illustration, 231 TA citations came from the USA, but only 115 of these were from articles that were originally authored by first-author affiliation from that territory, hence the other 116 were from other regions.

Table 52. Citations by author country

Count

OA			Region to region match		Total
			no match	match	
Toll Access	Region	USA	116	115	231
		Rest of Europe	104	16	120
		UK	48	12	60
		Rest of World	50	5	55
		China	70	20	90
		Pacific Rim	44	7	51
		Canada	26	1	27
		France	45	14	59
		Germany	59	11	70
		Italy	65	16	81
		Japan	28	24	52
		Spain	37	2	39
	Total		692	243	935
Open Access	Region	USA	251	359	610
		Rest of Europe	266	55	321
		UK	97	17	114
		Rest of World	99	5	104
		China	133	7	140
		Pacific Rim	114	10	124
		Canada	55	5	60
		France	108	29	137
		Germany	168	46	214
		Italy	93	27	120
		Japan	84	6	90
		Spain	60	3	63
	Total		1528	569	2097

Table 53 takes the data from

a step further and shows the citing country and the income group of the related cited articles. The 231 citations from the USA to the TA cited articles are shown by the World Bank per-capita income group from which they came, by first-author affiliation. It is evident that 115 of them were from the USA, as shown in

, but overall only 20 of the 231 were from regions outside the high per-capita income bracket. It is noticeable at this stage that a greater percentage of the TA cited articles (13.50%) are being cited outside of the high per-capita income bracket than OA articles, which account for only 4.7%.

Table 53 Citing country to cited income group

OA				Cited Income				Total
				Low	Lower middle	Upper middle	High	
Toll Access	Citing Country	USA	Count	0	17	3	211	231
			% within Region	.0%	7.4%	1.3%	91.3%	100.0%
		Rest of Europe	Count	0	7	10	103	120
			% within Region	.0%	5.8%	8.3%	85.8%	100.0%
		UK	Count	0	2	0	58	60
			% within Region	.0%	3.3%	.0%	96.7%	100.0%
		Rest of World	Count	0	3	9	43	55
			% within Region	.0%	5.5%	16.4%	78.2%	100.0%
		China	Count	3	21	6	60	90
			% within Region	3.3%	23.3%	6.7%	66.7%	100.0%
		Pacific Rim	Count	0	5	2	44	51
			% within Region	.0%	9.8%	3.9%	86.3%	100.0%
		Canada	Count	0	5	1	21	27
			% within Region	.0%	18.5%	3.7%	77.8%	100.0%
		France	Count	0	2	2	55	59
			% within Region	.0%	3.4%	3.4%	93.2%	100.0%
		Germany	Count	0	5	1	64	70
			% within Region	.0%	7.1%	1.4%	91.4%	100.0%
		Italy	Count	2	6	3	70	81
			% within Region	2.5%	7.4%	3.7%	86.4%	100.0%
		Japan	Count	0	3	1	48	52
			% within Region	.0%	5.8%	1.9%	92.3%	100.0%
		Spain	Count	0	3	4	32	39
			% within Region	.0%	7.7%	10.3%	82.1%	100.0%
	Total		Count	5	79	42	809	935
			% within Region	.5%	8.4%	4.5%	86.5%	100.0%
Open Access	Citing Country	USA	Count	0	14	12	584	610
			% within Region	.0%	2.3%	2.0%	95.7%	100.0%
		Rest of Europe	Count	0	11	7	303	321
			% within Region	.0%	3.4%	2.2%	94.4%	100.0%
		UK	Count	0	3	2	109	114
			% within Region	.0%	2.6%	1.8%	95.6%	100.0%
		Rest of World	Count	1	5	3	95	104
			% within Region	1.0%	4.8%	2.9%	91.3%	100.0%
		China	Count	0	7	4	129	140
			% within Region	.0%	5.0%	2.9%	92.1%	100.0%
		Pacific Rim	Count	0	12	0	112	124
			% within Region	.0%	9.7%	.0%	90.3%	100.0%
		Canada	Count	0	0	0	60	60
			% within Region	.0%	.0%	.0%	100.0%	100.0%
		France	Count	0	1	1	135	137
			% within Region	.0%	.7%	.7%	98.5%	100.0%
		Germany	Count	0	3	5	206	214
			% within Region	.0%	1.4%	2.3%	96.3%	100.0%
		Italy	Count	0	1	3	116	120
			% within Region	.0%	.8%	2.5%	96.7%	100.0%
		Japan	Count	1	1	1	87	90
			% within Region	1.1%	1.1%	1.1%	96.7%	100.0%
		Spain	Count	0	0	0	63	63
			% within Region	.0%	.0%	.0%	100.0%	100.0%
	Total		Count	2	58	38	1999	2097
			% within Region	.1%	2.8%	1.8%	95.3%	100.0%

The cluster bar chart in Figure 61 further extends the data from Table 53 by giving a comparative percentage of the whole count for each category by the OA status of the citations from each region. The USA, for example, receives 24.71% of all the citations to given to TA articles (231/935) and 29.09% of all those to given to OA articles (610/2097). Of the twelve regions, only five receive a greater overall percentage of the OA than of the TA citations, even though in every case each region had a greater number of OA than TA citations. It is noticeable that China receives 9.6% (90/935) of all citations given to TA articles but only 6.7% (140/2097) of all citations to OA articles. The Rest of World shows a similar

but narrower disparity: 5.9% (55/935) of all citations to TA articles but 5.0% (104/2097) of all those to OA articles.

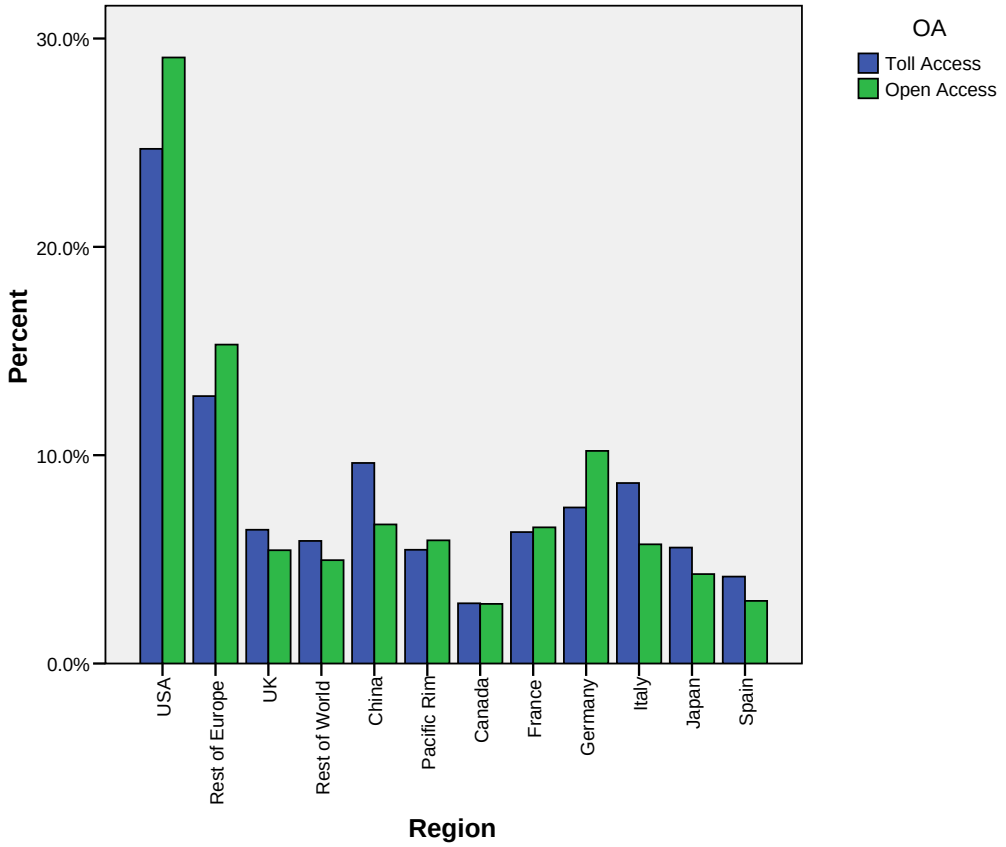


Figure 61 Percentage of citations to cited articles by OA status

Table 54 shows the distribution of citations to cited articles by their OA status and per-capita income group. What is evident is that there is a greater percentage of citations to the TA articles (20.00%, 187/935) from the low to upper middle income groups than is the case for OA articles, where the comparable group is 15.40% (323/2097) of all the citations to the OA articles.

Table 54. Cited to citing articles by income group

OA				Cited Country Income				Total
				Low	Lower middle	Upper middle	High	
Toll Access	Citing Country Income	Low	Count	0	1	0	5	6
			% within Citing Country Income	.0%	16.7%	.0%	83.3%	100.0%
		Lower middle	Count	3	22	8	73	106
			% within Citing Country Income	2.8%	20.8%	7.5%	68.9%	100.0%
		Upper middle	Count	0	6	15	54	75
			% within Citing Country Income	.0%	8.0%	20.0%	72.0%	100.0%
		High	Count	2	50	19	677	748
			% within Citing Country Income	.3%	6.7%	2.5%	90.5%	100.0%
	Total		Count	5	79	42	809	935
			% within Citing Country Income	.5%	8.4%	4.5%	86.5%	100.0%
Open Access	Citing Country Income	Low	Count	0	1	0	16	17
			% within Citing Country Income	.0%	5.9%	.0%	94.1%	100.0%
		Lower middle	Count	0	10	5	165	180
			% within Citing Country Income	.0%	5.6%	2.8%	91.7%	100.0%
		Upper middle	Count	1	3	6	116	126
			% within Citing Country Income	.8%	2.4%	4.8%	92.1%	100.0%
		High	Count	1	44	27	1702	1774
			% within Citing Country Income	.1%	2.5%	1.5%	95.9%	100.0%
	Total		Count	2	58	38	1999	2097
			% within Citing Country Income	.1%	2.8%	1.8%	95.3%	100.0%

Table 32 shows the distribution of the original 793 cited articles and the distribution of the 3032 citations by the country of their first-author affiliation. Countries are again classified by their World Bank per-capita income grouping. The number of articles appearing in each category has been given by its occurrence and the ratio is the division of the citing articles by the cited articles. The numbers in brackets indicate the number of article records in each category.

Table 32. Ratio of citing to cited articles by national income groups

Access Status	**World Bank classification by per capita income**			
TA Articles	Low	Lower middle	Upper middle	High
TA cited articles (298)	2	21	19	256
TA citing articles (935)	6	106	75	748
Ratio of citing to cited articles	3	5.05	3.95	2.92
OA Articles	Low	Lower middle	Upper Middle	High
OA cited articles (495)	1	14	18	462
OA citing articles (2097)	17	180	126	1774
Ratio of citing to cited articles	17	12.85	7	3.84

4. Results and Discussion

Overall there is a tendency for authors to cite work from their own country preferentially.

shows all the citations, analysed by whether there was a match or not between the nationality of the authors of the citing work and of the cited work. Generally, for every one citation that can be paired by country to the article it is citing, there are three that do not match. This applies to both OA and TA articles. Clearly, however, these data are skewed by the predominance of the region-to-region match for the USA. Given that just over 25% of the citations come from this territory alone, it is not surprising that of all citations almost 42% are to USA-affiliated first authors (data not shown). Perhaps this result is unremarkable, given that a large proportion (38.6%) of the cited articles originate from the USA, and that of all the citations from each region, the largest number are given to USA-affiliated authors. The 230 citations made by Chinese authors accounted for about 10% of all the citations to TA articles and about 7% of all the citations to OA articles. For the Pacific Rim and the Rest of the World territories, the overall TA/OA citation percentages were barely different, at around 5% each. This result appears to confirm the findings from the analysis by per capita income, that is, that there is little evidence to suggest that authors who live in countries that may have difficulties accessing TA journals are citing OA articles in greater numbers and hence boosting the citation count. Figure 61 shows that seven out of the twelve regions have more citations to TA than to OA articles. The seven regions include the UK, Italy, Japan, Spain, China and the Rest of the World, the latter two helping to support the premise that low-income does not generate exceptional OA citations. It is noticeable also, as demonstrated by her 116 were from other regions.

, that the regional link between cited and citing article by first author affiliation is generally weak, once the USA has been excluded. For citations of either access status, OA or TA, the overall regional match is about a quarter, but noticeably in the case of China, this is heavily skewed in favour of not citing other Chinese-affiliated authors.

Whilst there is unmistakable evidence from the data collected here that there is an overall citation advantage to those articles that are made available as OA (20%), the actual causes of this advantage, here and in other studies, are not always clear. Given that one of the primary arguments in favour of OA is that those who cannot afford access to peer-reviewed journal articles could use them if these articles were self-archived on the World Wide Web, where they could be readily accessed and cited by those with limited incomes. Hence, it could be reasoned that if this were clearly so, that a demonstrable cause of any citation advantage could be shown. Table 32 shows for the TA articles, the highest ratio of citing to cited articles occurs for citing authors in those countries in the lower middle income bracket, regardless of the nationality of origin of the cited articles. If all but the high income level countries are taken together, then the citation ratio is 4.45 for the TA articles and 9.79 for the OA articles. However, the overwhelming majority of articles are both authored and cited from the high-income countries. Table 32 gives the full data, divided into TA and OA articles, and by the four income categories of the first author's country of residence. Although this appears to be a convincing advantage, the percentage of the 187 lower-income citations to TA articles of all 935 citations to TA articles the result is 20%, and this is greater than the 15.40% figure from the comparable calculation for citations to OA articles. So there is a greater percentage of lower income authors among those citing TA articles than among those citing OA articles.

Given that self-citations have been eliminated from the original portion of the data, it is authors citing the work of others to support their work, who are doing the citing. As can be seen in Table 54, however, most of those who are doing the citing to the lower income groups are from the high-income countries. It is clear that a greater percentage of authors from the low and lower income countries cite more TA articles than OA articles, despite there being a higher ratio of citing to cited articles for those of OA status. In fact, 95.33% of all OA citations and 86.5% of TA citations are from high income regions.

5. Conclusions

Taken overall the results give a mixed picture as to whether those in the lower per capita income bracket countries are citing OA articles more frequently than TA ones.

The USA cites itself more than anyone else, which is not surprising given its level of authorship. The other developed countries, except for Japan, are all at about the same level in terms of within-nation citation. Table 32 suggests that while there is a modest difference between the citation ratios of OA and TA articles for citations given by authors in the developed world (3.84 versus 2.92), the difference becomes much greater when citations given by authors from the developing world are studied. The sample from the lowest income countries is very small. It may be that the lack of reliable telecommunications networks in these low income countries could hinder access to OA articles. In this case, scholars in these countries may rely on a limited number of printed journals for which they have subscriptions. The results from the larger sample in the lower middle income group of countries, however, are striking: a citation ratio of 12.85 for OA articles versus 5.05 for TA articles.

References

[1] Lawrence, S., 2001. Online or invisible. *Nature*, **411**(6837), 521.

[2] Antelman, K., 2004. Do open-access articles have a greater research impact? *College and Research Libraries*, **65**(5), 372-382.

[3] Eysenbach G., 2006. Citation advantage of open access articles. *PLoS Biology*, **4**(5), 692-698.

[4] Harnad, S. and Brody, T., 2004. Comparing the impact of OA (OA) vs. non-OA articles in the same journals. *D-Lib Magazine* [online], **10**(6), 1-5. <http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html>, [accessed 22.12.05].

[5] Kurtz, M. & Henneken, E., 2007. *Open access does not increase citations for research articles from The Astrophysical Journal.* <http://arxiv.org/ftp/arxiv/papers/0709/0709.0896.pdf>, [accessed 07.09.07].

[6] Moed, H., 2007. The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section. *Journal of the American Society for Information Science and Technology*. **58**(13), 2047-2054.

[7] Smith, J., 2007. Re: The apparent OA citation advantage. To multiple recipients of list *JISC Repositories*, 20 May, 19:19:35 BST.

[8] Norris, M., Oppenheim C. & Rowland, F. (in press). The citation advantage of open access articles. *Journal of the American Society for Information Science and Technology*.

HINARI, 2007. <http://www.who.int/hinari/en/>, [accessed 16.02.07].

AGORA, 2007. <http://www.aginternetwork.org/en/>, [accessed 16.2.07].

OAIster, 2007. < http://www.oaister.org/>, [accessed 07.03.07].

OpenDOAR. 2007. <http://www.OpenDOAR.org/search.php>, [accessed 07.02.07].

The World Bank, 2007. <http://go.worldbank.org/K2CKM78CC0>, [accessed 08.08.07].

Finding Open Access articles using Google, Google Scholar, OAIster and OpenDOAR

Online Information Review – *In press*

Michael Norris, Charles Oppenheim and Fytton Rowland
Department of Information Science, Loughborough University, UK

Abstract

Purpose – The paper seeks to demonstrate the relative effectiveness of a range of search tools in finding open access (OA) versions of peer reviewed academic articles on the WWW.

Design/methodology/approach – Some background is given to why and how academics may make their articles OA and how they may be found by others searching for them. Google, Google Scholar, OAIster and OpenDOAR were used to try to locate OA versions of peer reviewed journal articles drawn from three subjects (ecology, economics, and sociology).

Findings – Of the 2519 articles 967 were found to have OA versions on the WWW. Google and Google Scholar found 76.84% of them. The results from OpenDOAR and OAIster were disappointing, but some improvements are noted. Only in economics could OAIster and OpenDOAR be considered a relative success.

Originality/value The paper shows the relative effectiveness of the search tools in these three subjects. The results indicate that those wanting to find OA articles in these subjects, for the moment at least, should use the general search engines Google and Google Scholar first rather than OpenDOAR or OAIster.

Keywords: Open access; Google; Google Scholar; OpenDOAR; OAIster; Retrieval Performance

Paper type Research paper

Introduction

Academics can make their articles open access (OA) and thus freely available to anyone with Internet access by self-archiving electronic versions of their articles on their own personal web page, their department's web page, a subject repository or by depositing

them in an institutional repository as well as submitting them to an OA journal, a means of access not considered here. Articles deposited in repositories which use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) can have their metadata harvested, by for example, OAIster which currently holds the metadata of about 10 million articles. Anyone with Internet access may then search OAIster or use any of the general search engines to try and locate a particular author's work on the World Wide Web. It has been argued that articles, which are OA, accrue more citations than articles that remain behind subscription barriers (Harnad & Brody, 2004). Attempts to quantify this citation advantage have generally involved finding those articles that are OA on the World Wide Web and comparing their citation count to articles from the same journal issue, which remain accessible only by subscription. The mean citation counts of the two sets of articles are then compared (Antelman, 2004). In the first of two studies, similar in method to that conducted by Antelman the authors determined whether a particular set of OA articles did in fact have a citation advantage over their toll access (TA) counterparts (Norris, Oppenheim & Rowland in press). As part of the two studies, the authors used OAIster, OpenDOAR, Google and Google Scholar to try to locate as many OA versions of the articles as possible from the different subjects. The second study extended the first, by taking a further set of articles and used the same search tools to try to locate OA versions of them. This paper reports, primarily, the relative success of these search tools using article records from the second of the two studies.

Background

There are a growing number of institutional repositories that are OAI-PMH-compliant and consequently harvestable by service providers. Currently, the Registry of Open Access Repositories ROAR (2008) has over 1000 repositories registered worldwide, of which 536 are based at research institutions. These 536 archives hold a total of 2,309,512 records, averaging 5087 records each with a median figure of 938. In terms of the two million or so peer reviewed research articles published on a yearly basis, this represents a small but growing part of the total output. The graph shown in Figure 1 shows the rapid growth of institutional archives to March 2008 (Registry of Open Access Repositories (ROAR) 2008).

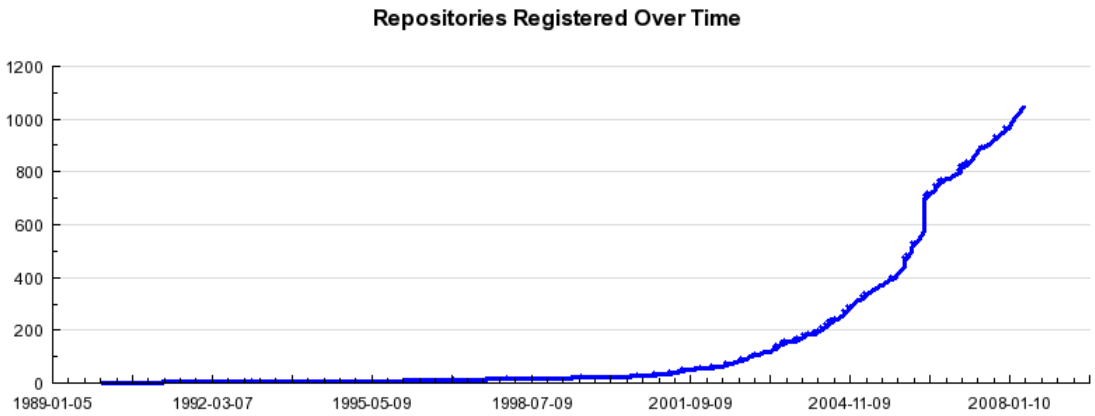


Figure 1. Repository numbers (ROAR 2008)

These repositories may be subject based like RePEC (Research Papers in Economics) or the physics archive arXiv, or may be more general in nature like the DEPOT or may be found at an institutional level. Authors may, of course self-archive their work to a home or departmental web page. OAIster, which harvests metadata is a union catalogue of digital sources hosted at the University of Michigan; "to establish a broad, generic retrieval service for information about publicly available digital library resources provided by the research library community" (About OAIster, 2007). Repositories make their records available to OAIster, where they harvest "their descriptive metadata (records) using OAI-PMH" (About OAIster, 2008). This service currently harvests from about 900 contributors and contains over 15 million records, which can be searched by author, title, language or subject and by resource type. Similarly OpenDOAR (About OpenDOAR, 2007), hosted by the University of Nottingham, facilitates access, worldwide, to institutional repositories. It is part of SHERPA (Securing a Hybrid Environment for Research Preservation and Access). Initially a directory of OA repositories, it now offers a trial service to search the contents of the repositories that it lists (OpenDOAR, 2006). Unlike OAIster, OpenDOAR does not search the repositories' metadata even if they are OAI-PMH compliant, but "relies on Google's indexes, which in turn rely on repositories being suitably structured and configured for the Googlebot web crawler" (SHERPA news, 2006). In contrast general search engines like Google and Google Scholar will search non-compliant home and departmental web pages as well as compliant OAI repositories.

Those undertaking research to see if there is an OA citation advantage have used differing strategies to find OA articles. Much research has focussed on the citation

advantage of articles deposited in arXiv that, on deposit, immediately become OA, compared to articles in the same subject that are not deposited and so remain TA (Kurtz *et al*. 2005: Moed 2007). Other authors have used various general search engines, either manually or by using robots to search the web for OA articles (Antelman 2004: Hajjem, Harnad, & Gringras (n.d.)). Whilst no one search engine can say that it indexes and searches all of the web, Google has established itself as the most frequently used general web search tool, with Google Scholar appearing as an addition to the Google stable of products focussing on scholarly materials. To a certain extent, Google and Google Scholar return many exclusive hits, returning differing results even when the same search terms are used.

In the case of Google and Google Scholar there have been many articles that have reviewed their performance and coverage. Google Scholar, in particular has had its coverage and citation structure reviewed by many authors. Jacsó has scrutinised closely the performance of Google Scholar, whilst in general highly critical, he does concede that "GS is good for locating relevant items, leading users some of the time to an open access version of a document, but it is not appropriate for bibliographic studies" (Jacso 2006, p. 307). Markland (2006) examined the effectiveness of both Google and Google Scholar at retrieving a defined set of items using keywords and title searches taken from 26 institutional repositories in the UK. Between them, using a title search, they found 25 out of the 26 items, with Google itself being the more successful finding all 25. Google Scholar found 17 items from within the repositories and found a further three items outside of the repositories. In contrast, when the repositories were searched directly using key words or titles taken from their own records, three items were not found. Walters compared the performance of Google Scholar to seven other databases (Academic Search Elite, AgeLine, AricleFirst, GEOBASE, POPLINE, Social Sciences Abstracts and Social Sciences Citation Index). He used a reference set of 155 articles on later life migration and found that Google Scholar found 93% of them, covering 27% more than Social Sciences Citation Index, its nearest rival.

Earlier work by the authors using Google Scholar in a pilot study carried out in late 2005, were disappointing, but subsequent work showed it to be more successful. It is assumed that in the intervening period between the pilot study and this research that the search capabilities of Google Scholar have been enhanced. This seems to be borne out

by recent comments from Jacsó (2008) who notes the increase in coverage of Google Scholar whilst still, however, deploring its software. On this basis, Google, Google Scholar, OAIster and the OpenDOAR service were used in combination as the search tools for finding OA versions of journal articles.

Methodology

A random sample of 628 articles was taken from the 10,119 that appeared in the 112 ecology journals listed in the 2005 Journal Citation Reports that were published in 2003. A purposive sample of 966 articles was taken from 21 mid-impact economics journals appearing in 2003 and 925 articles were taken from high impact sociology journals that appeared in 2004. The bibliographic details of each of the articles were taken. The four search tools, OAIster, OpenDOAR, Google and Google Scholar were identified as being likely to find as many OA articles as possible. The search for OA articles was conducted by entering the article's title as a phrase in each search tool. As the primary purpose of the research was to locate OA articles, the search sequence was designed to be progressive rather than exhaustive of each search tool, starting with OAIster, and then OpenDOAR, followed by Google Scholar, and finally Google. OAIster and OpenDOAR were always searched. If no hits were found using these two, then Google Scholar was searched; if Google Scholar also did not yield a result, then Google was also interrogated.

Results and discussion

Of the 2519 articles selected, 967 (38.39%) were found to have OA versions on the World Wide Web. Table 1 shows how these 967 articles are broken down by their OA status and subject.

Table 1. OA status by subject

Subject	Total Articles	% OA	% TA
Ecology	628	34.39	65.61
Economics	966	54.45	45.44
Sociology	925	24.32	75.68

Given the search protocol adopted, the results for Google and Google Scholar cannot be said to reflect the absolute potential of either of them. However, taken together, they

jointly found 76.84% of the articles. The percentage of records found for each search tool was; Google 8.79%, Google Scholar 68.04%, OAIster 2.38%, OpenDOAR 11.17% and where OAIster and OpenDOAR retrieved the same article, their combined score was 9.62%. Table 2 gives a more detailed breakdown of hits by subject and the search tool which found them. The hits for OAIster and OpenDOAR appear in columns four, five and six, columns four and five give exclusive hits and column six gives hits where both search tools have found the same record.

Table 2. Break down of OA hits by subject and search tool

Subject	Google	Google Scholar	OAIster	OpenDOAR	OAIster & OpenDOAR	Total
Ecology	20	194	2	0	0	216
Economics	32	287	13	108	86	526
Sociology	33	177	8	0	7	225
Total	85	658	23	108	93	967

Google Scholar was much more successful than OAIster and OpenDOAR, whose overall success was relatively poor. OAIster and OpenDOAR, however, could be considered useful search tools, finding 39.35% of the hits for economics. It is notable that in sociology, which had the smallest percentage of OA articles overall, that the majority of them were found using Google and Google Scholar, suggesting that those who do self-archive their work, in this subject at least, are not using repositories that can be found by using OAIster and OpenDOAR.

When the OA articles were broken down by first author affiliation, North America was found to be the region from which most articles originated. By subject, the percentage article counts from North America were, ecology 61.97%, economics 70.37% and for sociology 77.53% (data not shown).

There are major variations in OA hits when broken down by the search tool which found them and by first author affiliation. Table 3 shows the percentage of hits by each search tool. North America had the highest percentage of hits using Google and Google Scholar with the UK having the lowest percentage.

Table 3. Percentage OA hits by region and search tool.

Search tool	N America	Europe*	UK	Rest of World	Total
Google	4.34	2.59	0.93	0.93	8.79
Google Scholar	39.09	14.37	5.89	8.69	68.04
OAIster	1.24	0.52	0.41	0.21	2.38
OpenDOAR	4.76	3.62	1.65	1.14	11.17
OAIster & OpenDOAR	4.76	2.59	0.83	1.45	9.62
Total	54.19	23.68	9.72	12.41	100.00

* Does not include the UK

When OA hits were further examined by subject, their ranking by combined Google and Google Scholar were ecology 99.07%, economics 60.65%, and for sociology 93.33%. Given that both OAIster and OpenDOAR list the economics database RePEc among the sources from which they collect articles records, it is not surprising that there was a reasonable number of hits in this subject when using these two search tools. It is notable that OpenDOAR is overall more successful than OAIster in finding OA economics articles, presumably because it is searching RePEc, other repositories which allow Google's robots access.

As part of the first of the two studies undertaken by Norris, Oppenheim and Rowland (in press) they also examined, and briefly reported the success of the four search tools to find OA articles for the same subjects (including mathematics). Articles for this first study were taken from high impact journals from 2003. When the first study data is compared to the second, there are some notable differences. The second sample of high impact sociology articles were taken from 2004 and the percentage of hits dropped for *Google Scholar*, but rose for *Google* between the first and second study. Overall, their combined share of the hits for sociology drops from 98.37% to 93.33% to the benefit of OAIster. This is in contrast to the hits in ecology, where the combined *Google* score was 96.27% rising to 99.07% for the second study. Given that institutional repositories are more likely to be found at the more successful institutions (Directory of World Repositories 2008) and that the sample for the second round ecology data was randomly taken and hence more likely to come from a range of different institutions, it could be argued that it is more likely that the authors would self-archive to their own websites if repositories were not available at their own institution. However, for economics, where OpenDOAR was particularly successful, the combined scores for *Google* drops from

78.76% to 60.65% giving 39.35% of the share of the hits to OAIster and OpenDOAR in the second study, perhaps mirroring the growing success of these harvesters.

The relative success of OAIster and OpenDOAR is attributed to their harvesting the metadata from RePEc and the need for academics to share informal research results in general symposia and in working paper series. Antelman (2006, p.89) examined self-archiving practices within the social sciences, taking approximately 2000 articles from 22 high impact journals from 11 different publishers with varying self-archiving policies, including economics and sociology. For economics, she found an overall rate of self-archiving of 59% and for sociology 24%. For the two samples taken here, the rate for the economics' first study data was in the order of 65% and for the second round data 54.45% and for sociology 21% and 24.32% respectively, a noticeably similar result. Antelman goes on to explain the overall level of self-archiving as characteristic of the discipline, for the social sciences this is one where authors are less reliant on a culture of sharing information for example in the exchange of preprints. Economics, however, is characterised as a discipline with a higher degree of mutual dependence where working papers are shared through repositories with other authors. Apart from the RePEc there are few disciplinary depositories for the social sciences. Hence, there is little difference between the results between the first and second round studies for sociology, with the OA hits being found almost exclusively by *Google* and *Google Scholar* and with few academics archiving to any sort of repository.

Bergstrom and Lavaty (2007) used Google, Google Scholar and OAIster equally to help them find OA articles in economics and political science. From a sample of 703 economics articles, they could find most OA articles using Google, with Google Scholar finding some ten-percentage points less than Google. They found, using OAIster about 25% of their sample articles. This is a similar result to those found here in the second study, where 18.82% of the articles were located by searching OAIster. RePEc provided 27% of the articles, which is in itself, an interesting result given that OAIster lists RePEc as one of the sources it trawls. When the holdings of the two sources are compared, it is clear that not all the records available from RePEc are reported by OAIster and presumably, this explains the difference, although it is very unlikely that there were any items discovered in RePEc that could not be found by using Google or Google Scholar.

Conclusion

Despite the increasing number of institutional repositories and their harvesting by such services as OAIster, it is apparent that finding OA articles in the four subjects selected here was greatly facilitated by the use Google and Google Scholar. What is clear is that whilst OAIster and OpenDOAR are reliant for the majority of their content from institutional repositories, it appears that the majority of authors in this sample at least are not self-archiving their work to them or if they do, it is to non-compliant or unregistered repositories or to locations not accessible to these search tools. Alternatively, there may be of lack coverage by OAIster and OpenDOAR for other as yet unidentified reasons. Authors prefer, it seems, when they do self-archive their work, to do so to their personal or departmental web page where metadata harvesters such as OAIster cannot readily find them, but where Google and Google Scholar can. Those wanting to find OA articles, it is suggested, are more likely to find them using Google or Google Scholar rather than OpenDOAR or OAIster.

References

About OAIster. (2008), Available at: http://*OAIster*.umdl.umich.edu/o/*OAIster*/about.html

About OpenDOAR. (2006), Available at: http://*OpenDOAR*.org/about.html

Antelman, K. (2004), "Do open-access articles have a greater research impact", *College and Research Libraries*, Vol. 65 No 5, pp. 372-382.

Antelman, K. (2006) "Self-archiving practice and the influence of publisher policies in the social sciences", *Learned Publishing*, Vol. 19 No 2, pp. 85-95.

Bergstrom, T. and Lavaty, R. (2007) "How often do economists self-archive?", available at: http://repositories.cdlib.org/ucsbecon/bergstrom/2007a/

Directory of World Repositories. (2008), available at: http://www.webometrics.info/premierleague.asp

Hajjem, C., Harnad, S. and Gringras, Y., (n.d.), "Ten-year cross-disciplinary comparison of the growth of OA and how it increases citation impact", available at: http://eprints.ecs.soton.ac.uk/12906/

Harnad, S. and Brody, T. (2004), "Comparing the impact of OA (OA) vs. non-OA articles in the same journals", *D-Lib Magazine*, Vol. 10 No 6, pp.1-5. available at: http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/june04/harnad/06harnad.html

Jacso, P. (2006), "Deflated, inflated and phantom citation counts", *Online Information Review*, Vol. 30 No 3, pp. 297-309.

Jacso, P. (2008), "Savvy searching Google Scholar revisited", *Online Information Review*, Vol. 32 No 1, pp. 102-114.

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E and Murray, S. (2005), "The effect of use and access on citations", *Information Processing and Management*, Vol. 41 No 6, pp. 1395-1402.

Markland M., (2006), "Institutional repositories in the UK: what can the Google user find there?, *Journal of Librarianship and Information Science*, Vol.38 No 4. pp. 221-228.

Moed, H., (2007), "The effect of 'Open Access' upon citation impact: An analysis of ArXiv's Condensed Matter Section", *Journal of the American Society for Information Science and Technolog*, Vol. 58 No 13, pp. 2047-2054.

Norris, M., Oppenheim, C. and Rowland F. (In Press), "The Citation Advantage of Open Access Articles" *Journal of the American Society for Information Science and Technology*.

Registry of open access repositories (ROAR), (2008), available at: http://roar.eprints.org/index.php?action=generate_chart

SHERPA News. (2006), available at: http://www.sherpa.ac.uk/news/opendoaroct06.html

Walters, W. (2006), "Google Scholar coverage of a multidisciplinary field" *Information Processing and Management*, Vol. 43 No 4 pp. 1121-1132.

Presentation at the 12th Nordic Workshop 13-14th Sept 2007.

The Citation Advantage of Open Access Articles

The Citation Advantage of Open Access Articles

Michael Norris

Supervisors: Prof. Charles Oppenheim
& Dr Fytton Rowland

Introduction

- What are Open/Toll Access articles?
- The citation advantage
- Method
- Background data
- Findings
- Conclusion.

Open Access vs. Toll Access

- Open access articles are freely available on the Internet for anyone to read
- Toll access articles are only available by subscription
- OA articles are self-archived by their author to a web page or some type of repository e.g arXiv or RePEc from where they can be found or harvested.

Citation advantage

- That OA articles get more citations than TA articles
- That this advantage is measurable and evident for different subjects
- Earlier work – Lawrence, Harnad & Antelman.

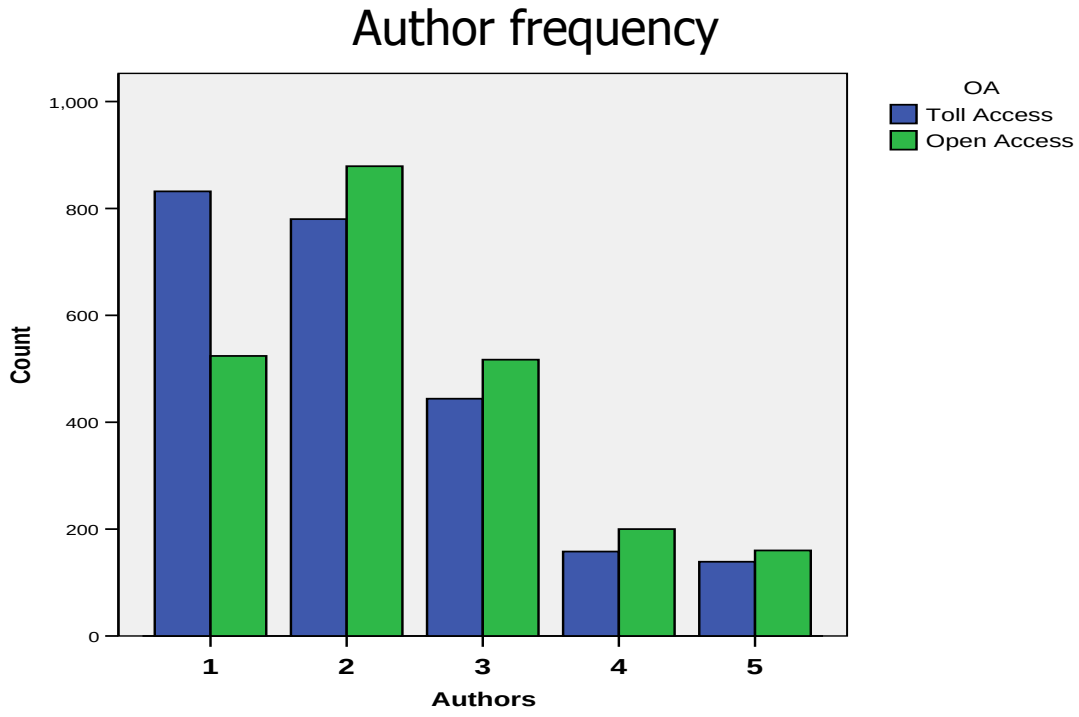
Method

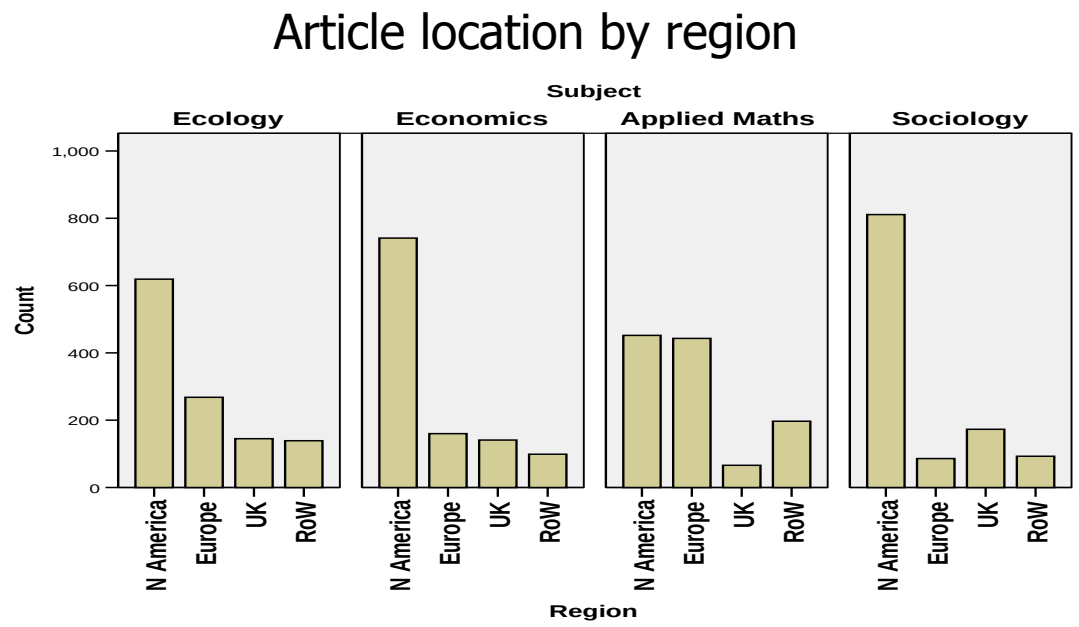
- Four subjects:-

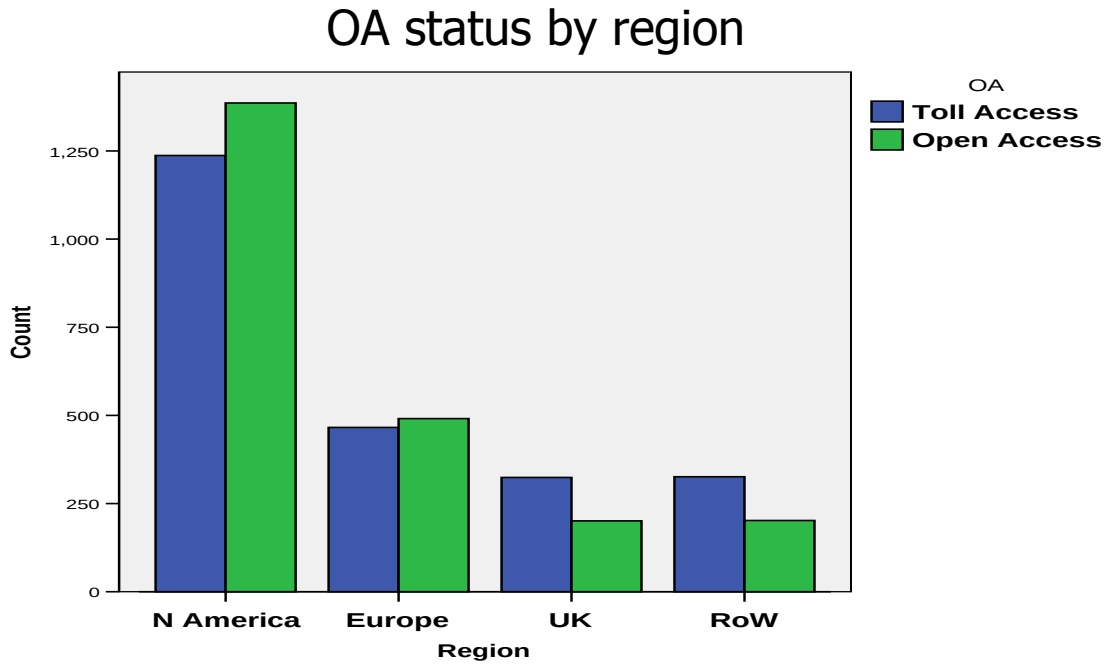
- Sociology
- Applied mathematics
- Ecology
- Economics.
- Articles taken from journals in 2003

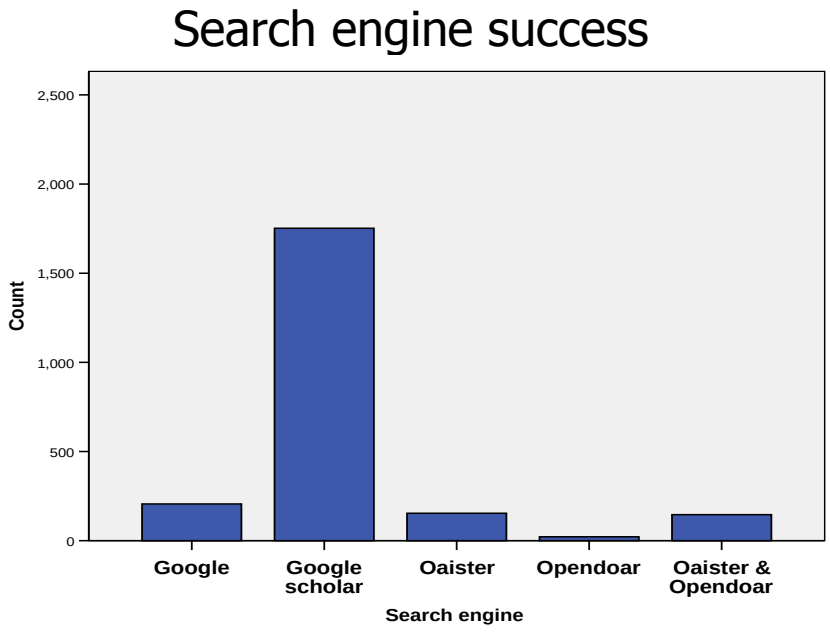
Method

- Citations counted – Web of Science
- Advantage = OA-TA/TA *100 citations
- Self, journal and other author citations were identified
- OA status found by searching Google, Google Scholar, OAIster and OpenDOAR

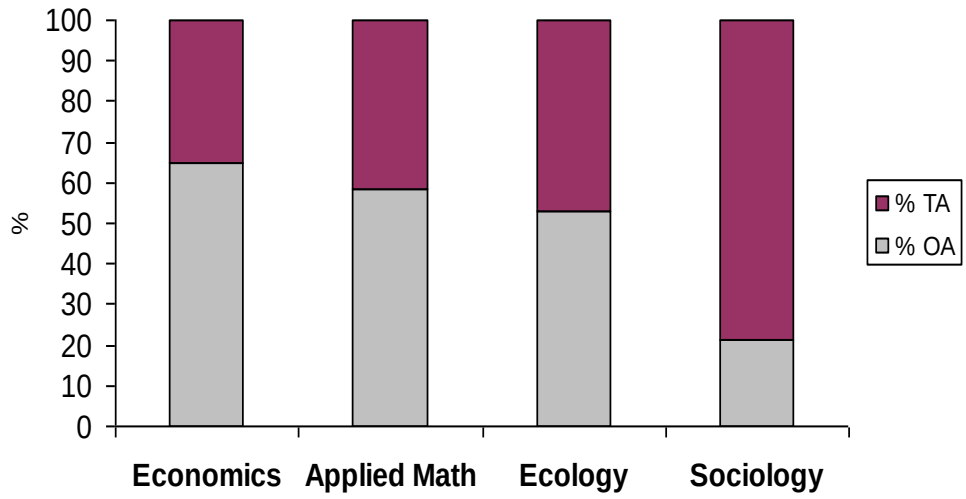

Author frequency
OA
Toll Access
Open Access
Count
1,000
800
600
400
200
0
1
2
3
4
5
Authors


Article location by region
Subject
Ecology
Economics
Applied Maths
Sociology
Count
1,000
800
600
400
200
0
N America
Europe
UK
RoW
Region


OA status by region
Count
1,250
1,000
750
500
250
0
N America
Europe
UK
RoW
Region
OA
Toll Access
Open Access


Search engine success
Count
2,500
2,000
1,500
1,000
500
0
Google
Google scholar
Oaister
Opendoar
Oaister & Opendoar
Search engine

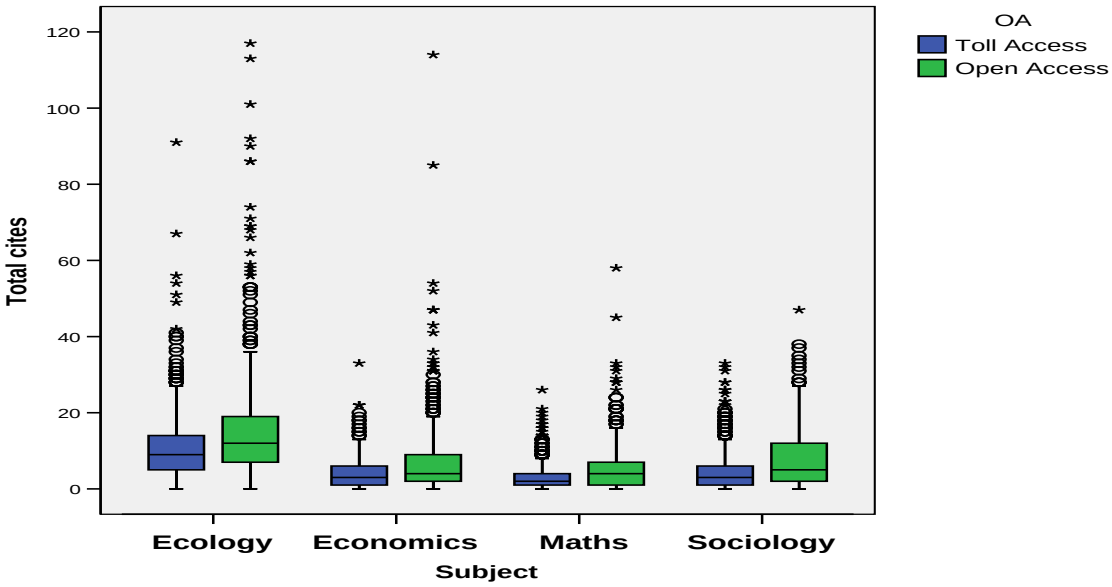
Findings
OA frequency



Findings
Citation advantage - %

	All citations	No self-citations
Sociology	88	103
Economics	62	77
Applied maths	53	71
Ecology	44	49

Findings
Citation advantage



Findings
Correlations

- Correlations are weak between:-

- Number of authors and citation counts;
- Journal impact factor and number of authors.

Findings
Associations

- But there are associations between:-

- Number of authors and OA status
- Journal impact factor and OA status
- Number of citations and OA status

Causes of the OA advantage

- Simply that the article is OA
- That it is available earlier
- That 'better' authors make their work OA
- That 'good' articles are made OA.

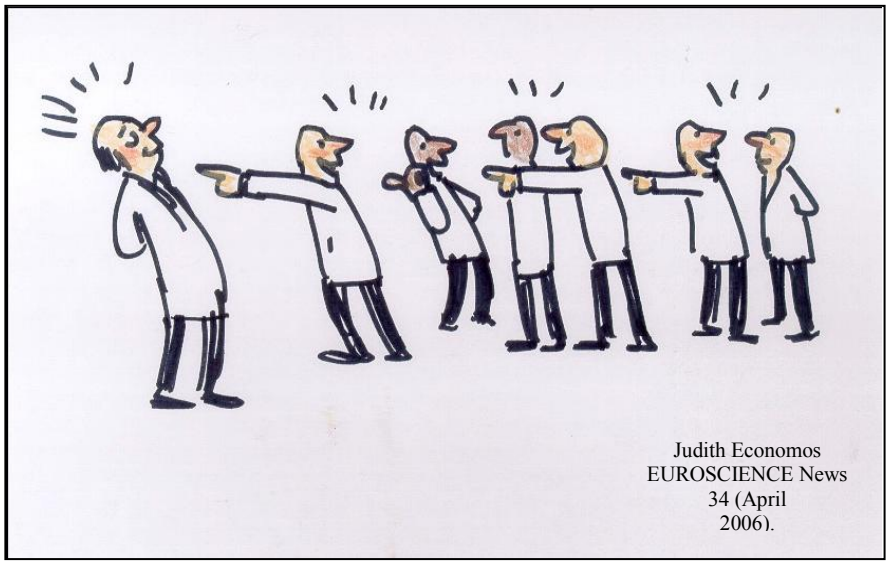
Conclusions

- There is a clear citation advantage

- Causes are less clear but:-

- Higher impact journals tend to have more OA articles
- The more authors the more likely an article will be OA
- The more citations an article has the more likely it will be OA

More OA citations might make you famous



Judith Economos
EUROSCIENCE News
34 (April
2006).

Thank you
Any Questions?