

## University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Evaluating Research Impact through  
Open Access to Scholarly Communication

by

Timothy David Brody

Thesis for the degree of Doctor of Philosophy

May 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS  
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

EVALUATING RESEARCH IMPACT THROUGH  
OPEN ACCESS TO SCHOLARLY COMMUNICATION

by Timothy David Brody

Scientific research is a competitive business – in order to secure funding, promotion and tenure researchers must demonstrate their work has impact in their field. To maximise impact researchers undertake high priority research, aim to get results first, and publish in the highest impact journals. The Internet now presents a new opportunity to the scholarly author seeking higher impact: s/he can now make their work instantly accessible on the Web through author self-archiving. This growing body of open access literature (coupled with new publishing models that make journals available for-free to the reader) maximises research impact by maximising the number of people who can read it, and making it available sooner. Open access also provides a new opportunity for bibliometric research. This thesis describes the relatively recent phenomenon of open access to research literature, tools that were built to collect and analyse that literature, and the results of analyses of the effect of open access and its effect on author behaviour. It shows that articles self-archived by authors receive between 50-250% more citations, that rapid pre-printing on the Web has dramatically reduced the peak citation rate from over a year to virtually instant and how citation-impact – now widely used for evaluation – can be expanded to include a new web metric of download impact.

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>                                      | <b>1</b>  |
| 1.1      | Motivation . . . . .                                     | 1         |
| 1.2      | Research Problem . . . . .                               | 4         |
| 1.3      | Approach . . . . .                                       | 6         |
| 1.4      | Terminology . . . . .                                    | 8         |
| 1.4.1    | arXiv Subject Headings . . . . .                         | 10        |
| 1.5      | Thesis Structure . . . . .                               | 11        |
| <b>2</b> | <b>The Scholarly Literature and Open Access</b>          | <b>13</b> |
| 2.1      | Background . . . . .                                     | 13        |
| 2.2      | Scholarly Publishing in Transition . . . . .             | 17        |
| 2.3      | The ‘Serials Crisis’ . . . . .                           | 19        |
| 2.4      | Open Access to Research Papers . . . . .                 | 20        |
| 2.4.1    | Author Self-Archiving . . . . .                          | 20        |
| 2.4.2    | Institutional Repositories . . . . .                     | 21        |
| 2.4.3    | Open Access Journals . . . . .                           | 23        |
| 2.5      | Policy Makers and Open Access . . . . .                  | 24        |
| 2.6      | Enabling Open Access using the Web . . . . .             | 26        |
| 2.7      | Evaluating Research Performance in Open Access . . . . . | 29        |
| 2.8      | Conclusion . . . . .                                     | 30        |
| <b>3</b> | <b>Citation Data and Analysis Tools</b>                  | <b>31</b> |
| 3.1      | Introduction . . . . .                                   | 31        |
| 3.2      | Commercial Services . . . . .                            | 31        |
| 3.2.1    | The ISI Web of Science . . . . .                         | 32        |
| 3.2.2    | The ACM Digital Library . . . . .                        | 34        |

|          |   |           |
|----------|---|-----------|
| 3.2.3    | Elsevier's ScienceDirect . . . . .                            | 34        |
| 3.3      | Open Access to Citation Data . . . . .                        | 35        |
| 3.3.1    | RePEc . . . . .   | 36        |
| 3.3.2    | Citeseer . . . . .  | 37        |
| 3.3.3    | The Open Citation Project . . . . .                           | 38        |
| 3.4      | The Open Archives Initiative . . . . .                        | 39        |
| 3.4.1    | Dublin Core . . . . .   | 41        |
| 3.4.2    | Metadata Semantic Problems . . . . .                          | 43        |
| 3.5      | OpenURL . . . . .   | 44        |
| 3.5.1    | Persistent Linking using OpenURL . . . . .                    | 44        |
| 3.5.2    | Contextual Linking using OpenURL and SFX . . . . .            | 45        |
| 3.5.3    | Digital Object Identifier . . . . .                           | 47        |
| 3.6      | Conclusion . . . . .  | 48        |
| <b>4</b> | <b>Bibliometrics</b>  | <b>50</b> |
| 4.1      | Bibliometric Techniques and Laws . . . . .                    | 51        |
| 4.1.1    | Zipf's Law . . . . .  | 51        |
| 4.1.2    | Lotka's Law . . . . .   | 51        |
| 4.1.3    | Bradford's Law . . . . .                                      | 53        |
| 4.1.4    | Bibliographic Coupling and Co-Citation . . . . .              | 53        |
| 4.2      | Eugene Garfield and the Science Citation Index . . . . .      | 54        |
| 4.2.1    | The Impact Factor . . . . .                                   | 55        |
| 4.3      | Conclusion . . . . .  | 57        |
| <b>5</b> | <b>An Analysis of OAI Repositories and Harvesting Support</b> | <b>58</b> |
| 5.1      | Introduction . . . . .  | 58        |
| 5.2      | Celestial Architecture . . . . .                              | 59        |
| 5.3      | Reducing Repeated Requests . . . . .                          | 61        |
| 5.4      | Abstracting Multiple OAI Protocol Versions . . . . .          | 62        |
| 5.5      | Correcting OAI Data Provider Errors . . . . .                 | 63        |
| 5.6      | Analysing OAI Data Providers . . . . .                        | 66        |
| 5.7      | Adding Repositories to Celestial . . . . .                    | 68        |
| 5.8      | Conclusion . . . . .  | 69        |
| <b>6</b> | <b>Quantifying Open Access in Institutional Repositories</b>  | <b>71</b> |

|          |   |           |
|----------|---|-----------|
| 6.1      | Introduction . . . . .  | 71        |
| 6.1.1    | Repositories <i>vs.</i> Archives . . . . .  | 72        |
| 6.1.2    | Existing and New Lists of Repositories . . . . .  | 74        |
| 6.2      | The Registry of Open Access Repositories . . . . .  | 76        |
| 6.2.1    | Criteria for Inclusion in ROAR . . . . .  | 76        |
| 6.2.2    | Removal from the Registry . . . . .   | 78        |
| 6.2.3    | Adding Records to the Registry . . . . .  | 78        |
| 6.2.4    | Maintaining the Registry . . . . .  | 79        |
| 6.2.5    | Creating Web Page Thumbnails . . . . .  | 81        |
| 6.3      | Analysis of Institutional Repositories . . . . .  | 83        |
| 6.3.1    | Software Installations . . . . .  | 83        |
| 6.3.2    | Countries . . . . .   | 88        |
| 6.4      | Peer-reviewed Full-Text Detection . . . . .   | 89        |
| 6.4.1    | Full-text Detection Implementation . . . . .  | 90        |
| 6.4.2    | Full-text Results . . . . .   | 91        |
| 6.5      | Conclusion . . . . .  | 92        |
| <b>7</b> | <b>Building an Open Access Citation Index</b>   | <b>93</b> |
| 7.1      | Introduction . . . . .  | 93        |
| 7.2      | Citebase Search and the OAI-PMH . . . . .   | 95        |
| 7.3      | Reference Parsing . . . . .   | 96        |
| 7.3.1    | TeX . . . . .   | 96        |
| 7.3.2    | XML . . . . .   | 97        |
| 7.3.3    | Unstructured Documents . . . . .  | 98        |
| 7.3.4    | Citation linking . . . . .  | 99        |
| 7.4      | Citebase Search's Web Interface . . . . .   | 100       |
| 7.4.1    | Ranking Search Results by Citation Impact . . . . .   | 104       |
| 7.4.2    | The Hub/Authority Algorithm . . . . .   | 105       |
| 7.5      | Citebase Search Database Structure . . . . .  | 106       |
| 7.6      | Citebase Search Analysis Tools . . . . .  | 107       |
| 7.6.1    | Paper Frequency: <code>article_frequency</code> , <code>article_frequency_ads</code> ,<br><code>papers_per_field</code> . . . . .               | 108       |
| 7.6.2    | Citation Histogram: <code>citation_frequency</code> , <code>citation_zipf</code> . . . . .  | 109       |
| 7.6.3    | Citation Latency: <code>citation_latency</code> , <code>reference_latency</code> ( <code>per_area</code> ,<br><code>per_year</code> ) . . . . . | 109       |

|          |   |            |
|----------|---|------------|
| 7.6.4    | Age of Cited Papers: cited_age . . . . .  | 110        |
| 7.6.5    | Downloads Analysis: hits_frequency, hits_latency, hits-<br>bydomain . . . . .                       | 111        |
| 7.6.6    | Downloads-Citations Latency Comparison: hits_latency_normalised,<br>hitslatencybyquartile . . . . . | 112        |
| 7.6.7    | The Correlation Generator . . . . .   | 113        |
| 7.7      | Citebase Search Usage Analysis . . . . .  | 113        |
| 7.8      | Conclusion . . . . .  | 114        |
| <b>8</b> | <b>Using Web Statistics for Usage Analysis</b>  | <b>116</b> |
| 8.1      | Introduction . . . . .  | 116        |
| 8.1.1    | Why predict citation counts? . . . . .  | 117        |
| 8.1.2    | Web accesses as an early-day predictor . . . . .  | 118        |
| 8.2      | Chapter Structure . . . . .   | 118        |
| 8.3      | arXiv . . . . .   | 118        |
| 8.4      | Harvesting from arXiv . . . . .   | 121        |
| 8.5      | Citebase . . . . .  | 122        |
| 8.6      | Accuracy of Citation Links within Citebase . . . . .  | 123        |
| 8.7      | Correlation between Citations and Downloads . . . . .   | 126        |
| 8.8      | Correlation Generator . . . . .   | 127        |
| 8.9      | Correlation Generator Implementation . . . . .  | 130        |
| 8.10     | Generating Correlations . . . . .   | 133        |
| 8.11     | Sample Correlations . . . . .   | 134        |
| 8.12     | Predicting citation impact from downloads . . . . .   | 136        |
| 8.13     | Predicting citation impact from early-day citations . . . . .                                       | 138        |
| 8.14     | Conclusion . . . . .  | 140        |
| <b>9</b> | <b>The Effect of Open Access on Citation Behaviour</b>  | <b>141</b> |
| 9.1      | Introduction . . . . .  | 141        |
| 9.2      | Increased Citation Impact due to Open Access . . . . .  | 142        |
| 9.2.1    | Methodology . . . . .   | 143        |
| 9.2.2    | Results . . . . .   | 144        |
| 9.3      | Reduced Citation Latency due to Pre-Print Archiving . . . . .                                       | 151        |
| 9.3.1    | Decreasing e-print Citation Latency . . . . .   | 152        |
| 9.4      | Citation Obsolescence . . . . .   | 153        |

|   |            |
|---|------------|
| 9.5 Conclusion . . . . .  | 156        |
| <b>10 Conclusions and Future Directions</b>   | <b>158</b> |
| 10.1 Introduction . . . . .   | 158        |
| 10.2 Contribution of this Thesis . . . . .  | 159        |
| 10.3 Open Access Improves Citation Impact and Decreases Citation<br>Latency . . . . . | 160        |
| 10.4 Web Impact as a Predictor of Citation Impact . . . . .                           | 161        |
| 10.5 Maximising the Benefit of Research Funding . . . . .                             | 162        |
| 10.6 Future Directions . . . . .  | 163        |
| 10.6.1 Extending and supporting this work . . . . .                                   | 164        |
| 10.7 In Summary . . . . .   | 165        |
| <b>Bibliography</b>   | <b>166</b> |



# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | “And I can’t tell you the rest until the journal comes out.”<br>Bachrach et al. (1998)                            | 1  |
| 3.1 | ‘Cited References’ in the ISI Web of Science  | 33 |
| 3.2 | ‘Citing Articles’ in the ISI Web of Science   | 33 |
| 3.3 | Citation links in the ACM Digital Library.  | 35 |
| 3.4 | References are linked where the cited item is known to ScienceDirect.   | 36 |
| 3.5 | Following the ‘Cited By’ link shows a list of papers citing the current paper.                                    | 36 |
| 3.6 | Liu et al. (2002) proposed infrastructure for extending OAI-PMH   | 39 |
| 3.7 | High-level OAI-PMH data flow-chart.   | 40 |
| 4.1 | Papers rank-ordered by citation impact and a Zipfian distribution (double-logarithmic axis)                       | 52 |
| 4.2 | Lotka’s Law distribution (1 contribution = 100)   | 52 |
| 5.1 | Celestial’s Architecture  | 60 |
| 5.2 | Celestial supports experimental services by avoiding the need to repeatedly harvest source repositories.          | 62 |
| 5.3 | A bad character gets ignored by Internet Explorer when rendering a web page                                       | 64 |
| 5.4 | In XML a bad character prevents the document from being parsed  | 64 |
| 5.5 | Thumbnail graph of arXiv’s records generated by Celestial for ROAR, that shows two jumps in the number of records | 67 |
| 5.6 | Celestial summary graph for Bepress-based repositories  | 68 |
| 5.7 | Adding repositories to Celestial by URL   | 69 |

|      |  |     |
|------|--|-----|
| 5.8  | Selecting which URLs to add . . . . .  | 69  |
| 5.9  | The URL is recognised as an OAI-PMH interface and ready<br>for harvesting . . . . .  | 69  |
| 6.1  | ROAR metadata entry form with embedded preview. . . . .  | 79  |
| 6.2  | Updating an entry in ROAR . . . . .  | 81  |
| 6.3  | Minor customisations of DSpace . . . . .   | 81  |
| 6.4  | Major customisations of DSpace . . . . .   | 82  |
| 6.5  | ROAR country-breakdown . . . . .   | 83  |
| 6.6  | Rate of growth of GNU EPrints-based repositories and contents  | 84  |
| 6.7  | Rate of growth of DSpace-based repositories and contents . . .   | 84  |
| 6.8  | ECS EPrints contents, based on its ‘Advanced Search’ . . . . .   | 86  |
| 7.1  | Searching Citebase Search by Metadata query. . . . .   | 94  |
| 7.2  | Compound reference. . . . .  | 96  |
| 7.3  | XML format reference. . . . .  | 97  |
| 7.4  | Total references parsed and linked in Citebase Search, per month.  | 100 |
| 7.5  | Results from a metadata search for “Witten, E” and “String<br>Theory”. . . . .   | 101 |
| 7.6  | Citebase Search abstract page (jagged line is abbreviation). . .   | 101 |
| 7.7  | Download and citation summary table and figure. . . . .  | 102 |
| 7.8  | Downloads by country. . . . .  | 102 |
| 7.9  | Downloads over time. . . . .   | 103 |
| 7.10 | Citebase Search reference list. . . . .  | 103 |
| 7.11 | Metadata summary (taken from ‘Top 5 citing papers’). . . . .   | 103 |
| 7.12 | Bibliographically coupled papers search. . . . .   | 104 |
| 7.13 | Citebase Search search result rankings. . . . .  | 104 |
| 7.14 | Hub-Authority algorithm (HITS). . . . .  | 106 |
| 7.15 | The <b>statistics</b> script provides links to all available analyses.   | 108 |
| 7.16 | Papers frequency in arXiv (left) and the ADS data set (right)  | 108 |
| 7.17 | Histogram of all papers by citation impact (left) and papers<br>rank-ordered by citation impact plotted on logarithmic scales<br>(right) . . . . . | 109 |
| 7.18 | Reference latency for all papers (left) and broken down by<br>arXiv subject (right) . . . . .  | 110 |

|      |  |     |
|------|--|-----|
| 7.19 | Citation latency per year (based on year of accession of the cited paper) . . . . .  | 110 |
| 7.20 | Plotting the cited age uses all references, because it only requires the year of publication (highlighted in red) . . . . .                                    | 111 |
| 7.21 | Year of publication of cited papers, by year of the citing arXiv paper . . . . .   | 111 |
| 7.22 | Histogram of arXiv papers deposited in 2000 by the number of downloads to each paper (left) and histogram of hits by age of downloaded paper (right) . . . . . | 112 |
| 7.23 | Histogram of papers by download impact, separated into quartiles by citation impact . . . . .  | 113 |
| 7.24 | Citebase Search usage in 2005 (excludes web crawlers and robots).114   |     |
| 8.1  | Monthly submission rate for arXiv (source arXiv) . . . . .   | 119 |
| 8.2  | Recent growth in arXiv is due to Cond-Mat and Astro-Ph . . . . .   | 120 |
| 8.3  | Age of papers downloaded or cited in 2004/09 (normalised). . . . .   | 127 |
| 8.4  | Example distribution with Pearson's $r$ line plotted . . . . .   | 128 |
| 8.5  | Correlation scatter graph without logarithmic translation . . . . .  | 129 |
| 8.6  | Correlation scatter graph with logarithmic translation . . . . .   | 130 |
| 8.7  | Preview graphs show the distribution of key variables and allow visual parameter selection . . . . .   | 131 |
| 8.8  | Available parameters for customising the correlation calculation   | 132 |
| 8.9  | Scatter-plot for all papers deposited in arXiv between 2000-2003 (excluding first seven days of downloads) . . . . .   | 134 |
| 8.10 | Scatter-plot for <i>hep</i> papers deposited in 2000-2003 . . . . .  | 135 |
| 8.11 | Scatter-plot for HEP papers deposited in 2000-2001, counting only citations and downloads up to two years after deposit . . . . .                              | 135 |
| 8.12 | Example download and citation windows used for prediction calculations. . . . .  | 136 |
| 8.13 | The predictive power of downloads reaches an asymptote at 6 months. . . . .  | 137 |
| 8.14 | Correlation between papers' citation impact at one month and two years. . . . .  | 138 |
| 8.15 | Correlation between papers' citation impact at six months and two years. . . . .   | 139 |

|      |  |     |
|------|--|-----|
| 9.1  | Subject selection form. . . . .  | 145 |
| 9.2  | Open access advantage for Nuclear & Particle Physics, controlled by by-subject . . . . .   | 146 |
| 9.3  | Open access advantage for Nuclear & Particle Physics, controlled by by-journal . . . . .   | 146 |
| 9.4  | Open access advantage for Nuclear & Particle Physics, controlled by by-journal and excluding self-citations . . . . .                          | 147 |
| 9.5  | Open access advantage for Nuclear & Particle Physics, controlled by by-journal, excluding self-citations and using same-size samples . . . . . | 147 |
| 9.6  | Open access advantage for all Physics fields. OAA*OAP $r = 0.441$ , OAP*Year $r = 0.953$ , OAA*Year $r = 0.624$ . . . . .                      | 149 |
| 9.7  | Open access advantage for General Physics. OAA*OAP $r = 0.489$ , OAP*Year $r = 0.983$ , OAA*Year $r = 0.408$ . . . . .                         | 149 |
| 9.8  | Open access advantage for Nuclear & Particle Physics. OAA*OAP $r = 0.092$ , OAP*Year $r = 0.848$ , OAA*Year $r = 0.153$ . . . . .              | 150 |
| 9.9  | Citation latency is the time delay between a paper $A$ being deposited and a citing paper $B$ being deposited . . . . .                        | 151 |
| 9.10 | Annual citation latency for arXiv . . . . .  | 153 |
| 9.11 | The age of papers cited by arXiv High Energy Physics papers  | 154 |
| 9.12 | The age of papers cited by arXiv Astronomy/Astrophysics papers   | 155 |
| 9.13 | The age of papers cited by arXiv Maths papers . . . . .  | 155 |

# List of Tables

|     |  |     |
|-----|--|-----|
| 1.1 | arXiv sub-archives . . . . .   | 10  |
| 3.1 | Dublin Core Metadata Element Set . . . . .   | 42  |
| 3.2 | Example KEV-encoded OpenURL . . . . .  | 45  |
| 5.1 | Top ten most widely implemented metadata formats in Celestial-<br>registered repositories . . . . .          | 61  |
| 5.2 | Character encoding in UTF-8 and latin-1 . . . . .  | 65  |
| 6.1 | ROAR metadata fields . . . . .   | 80  |
| 6.2 | Software types configured in ROAR . . . . .  | 85  |
| 6.3 | Top 11 countries ranked by number of repositories . . . . .  | 88  |
| 6.4 | Top 11 countries ranked by number of institutional records,<br>normalised by population . . . . .            | 89  |
| 6.5 | Number of OAI records pointing to full-texts . . . . .   | 91  |
| 7.1 | Downloads by country between 2005-07 and 2006-07 (1257768<br>total). . . . .                                 | 112 |
| 8.1 | Summary of sample set used to test Citebase's reference pars-<br>ing and citation linking accuracy . . . . . | 125 |
| 8.2 | Summary of reference links from an example paper. . . . .  | 125 |
| 8.3 | Most recent 90 papers . . . . .  | 125 |
| 9.1 | Cumulative percentage of cited literature by age for High En-<br>ergy Physics e-prints. . . . .              | 156 |

## DECLARATION OF AUTHORSHIP

I, **Timothy David Brody**

declare that the thesis entitled

### **Evaluating Research Impact through Open Access to Scholarly Communication**

and the work presented in are my own. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

Brody, T., Harnad, S. and Carr, L. (2005) Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology*. *In Press*

**Signed:** .....

**Date:** .....

## Acknowledgements

Thank you to my supervisors, Professor Stevan Harnad and Dr. Leslie Carr, who have inspired, encouraged and funded me through the many years this work has taken. Getting a PhD should be my ‘optimal and inevitable’, but – perhaps like all worthwhile things – it takes a while getting there. Thanks to Tim Miles-Board who provided invaluable advice on planning this thesis. My doctorate was funded by the Engineering and Physical Sciences Research Council (EPSRC).

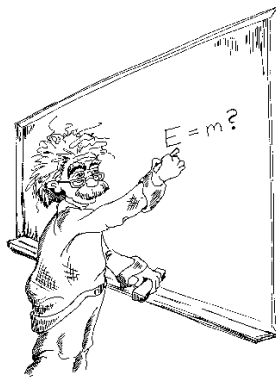
While working on a doctorate has sometimes felt like a solo endeavour, much of this work has relied on feedback from colleagues and users. In particular I would like to thank Steve Hitchcock, Chris Gutteridge and Jessie Hey who provided a sounding board and critical feedback.

I’ve had the opportunity to present my work at many international conferences and workshops, allowing me to meet fellow researchers from a wide range of countries. My thanks go to the conference organisers who invited me – it has greatly helped me.

For my parents – Pat and Simon.

# Chapter 1

## Introduction



**Figure 1.1:** “And I can’t tell you the rest until the journal comes out.” [Bachrach et al. \(1998\)](#)

### 1.1 Motivation

Generating performance indicators for research authors was not my goal when I first started working on bibliometrics. In the summer of 2000 I was contracted as an undergraduate intern to analyse data collected from the Open Citation Project ([Hitchcock et al., 2000](#), ran from 1999-2002). That analysis resulted in a series of questions and results collectively called “Mining the Social Life of an ePrint Archive” ([Brody et al., 2000](#)). The available data included web download logs and a small citation database, used to link references in PDF papers to the cited paper from the UK arXiv



e-print server (a collection of author-deposited research papers). This analysis took the form of posing a list of questions, *e.g.* “Is there a relation between the number of times a paper is downloaded and number of citations it receives?” Although much of that work was limited by what was possible with the data at hand and the time limit of a summer job, it is still the foundation for this PhD.

From humble beginnings with the Open Citation Project’s database my analysis of web logs and citation data has continued and expanded throughout the duration of my PhD (begun in 2001). The amount of raw data available has increased as more authors make their research results freely available through arXiv and other e-print servers, better known now as *open access* (see [Chan et al., 2002](#), the *Budapest Open Access Initiative*). The scope of my research has expanded to try to identify the consequences that open access has for scholarly communication.

As well as analysis I had built rudimentary web tools using OpCit’s data, including a citation navigation tool that allowed the user to interactively expand the citation tree, showing the citing papers of cited papers *etc.* The most complete of these tools was *Citebase Search* – a web-based search engine that ranked research papers by citation impact. What precipitated Citebase Search was the development of the *Open Archives Initiative*; the OpCit Project was a participant at OAI’s inaugural meeting at Santa Fe ([Van de Sompel and Lagoze, 2000](#)). arXiv was one of the test-beds for developing the OAI Protocol for Metadata Harvesting (OAI-PMH) which allows services to download metadata (descriptions of records) from OAI-compliant archives. By coupling a database of metadata downloaded from arXiv using the OAI-PMH with OpCit’s citation data Citebase Search provided a metadata-search engine that could rank-order arXiv papers by their citation count.

Citebase Search was integrated into the demonstration OpCit system by adding links into OpCit’s reference-linked PDFs. Although it was hoped that arXiv might incorporate OpCit’s linking directly into their service it was never adopted but, by developing a separate service, the reference linking developed by OpCit could be demonstrated. While arXiv showed

little interest in adopting reference linking internally, they have provided links to Citebase Search (for “autonomous citation navigation and analysis”<sup>1</sup>) which has greatly improved the visibility and use of the service.

The goal of rank-ordering search results by citation counts is to allow more significant papers to come higher in the search results. Citations are especially useful in search results, as a citation is – in effect – an *endorsement* by an author, hence the paper is more likely to be useful compared to papers that haven’t been cited. However, citation counts can also be used to *evaluate* research papers, the journals they’re published in, and the authors that wrote them. It is this quantitative evaluation that relates to the first part of this thesis; the claim that this paper is better than that paper because it has been cited more.

Citebase Search has grown steadily as the amount of literature made available by authors from arXiv and other archives has increased. In developing Citebase Search the potential benefit that open access has for service provision has become clear. As authors self-archive or publish in open access journals, so the value and potential of services built on that literature grows (the usefulness of a service is dependent on the amount of material contained *e.g.* whether the user can seamlessly follow citations depends on the cited paper being available). Because of open access Citebase Search hasn’t needed to negotiate licences for the content it contains, as all the source material is available to any service.

Services built on open access therefore compete not on the basis of how much content they contain, but on the quality of the service that they provide. This leads to a particularly user-driven environment – if a service doesn’t provide real value to the user, another provider can easily create a better, competing service with the same content.

If the benefits of open access can be proved that allows policy makers to mandate authors to make their work freely available on web, the result of which feeds into open access services that enable greater competition and innovation, which is of ultimate benefit to authors. Proving the benefits of

---

<sup>1</sup>See any arXiv abstract page, *e.g.* <http://arxiv.org/abs/astro-ph/0602632>

open access requires quantitative analysis of the effect that open access has on the research literature. I've used Citebase Search to monitor the effect that open access has had on the arXiv author community.

## 1.2 Research Problem

The justification for providing open access is that the current system of access prevents potential users of research literature from being able to access the material they need to perform their own research. Most research is published in journals or conference proceedings that are only accessible by paying a subscription fee and with 2.5 million research papers ([Harnad et al., 2004](#)) being published no research library can afford access to all the world's research output .

The level of access denied to users depends on where they are – ranging from the “Harvards to the Have-nots<sup>2</sup>”. Even Harvard – with the largest collection of subscription content in the world (see [Kyrillidou and Young, 2004](#), table 3) – will still be missing out because of the limited distribution of national publications. Despite publisher deals the situation for researchers in developing countries is much worse ([Chan and Kirsop, 2001](#); [Smith, 2004a](#)). Search and indexing services are also limited because they are subject to the same restrictions to subscription-based journals – forcing users to use many search tools in addition to being denied access to papers found by a search but to which they don't have a subscription.

The trouble for users is that they can do little about what is available to them, being limited by the funds available to purchase subscriptions. However, lack of reader access feeds back to authors through lost usage and impact. Authors write papers that cite material they have used to perform their research. The citation impact of a paper is therefore dependent upon authors (as readers) being able to discover, read and subsequently cite that paper. If the current model of access to research fails to provide all potential citing authors with access then there is lost usage and citations,

---

<sup>2</sup>Stevan Harnad, September '98 forum

hence authors don't maximise their potential research impact. As research authors publish their work primarily to gain recognition and promotion, rather than for direct financial gain through royalties, this reduces their and – by proxy – their institution's prospects.

[Hajjem et al. \(2005\)](#) found only 5%-16% of papers indexed by the *ISI Science Citation Index* were available as open access. If open access can be shown to maximise research impact (by lowering the bar of access to that of access to the web), then there is a strong empirical and economic argument to convince authors and institutions of the benefit of providing open access to their work. Of course, once all research material is openly accessible, then there would be no competitive benefit for authors that provide open access versus authors accessible only through subscription-based access. But by then all researchers would see the benefit of having instant access to all the world's research.

Open access research material can be roughly separated into four sources: open access journals, subject-based research repositories, institutional repositories and personal web sites. A scholarly search tool would ideally allow the user to search or navigate across all open access material, providing the most useful matches first. An approach to providing this ability is to retrieve data from open access sources into a central, aggregated service that provides a search across all of the harvested content. For open access journals and repositories the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides a means of transferring this data from the repository to the service.

The difficulty with providing an aggregated service is that the repositories harvested from are heterogeneous; repositories will collect different amounts of data from depositors, and may not provide the same meaning to a particular metadata field. Not all of the information required to provide a federated service may be available from the repository's OAI-PMH interface, in which case it may be necessary to perform separate web queries to the web interface of the repository, *e.g.* to retrieve the full text.

In almost all situations an aggregating service will need to perform some

form of normalisation and verification of incoming data. Recent web-based services have shown that much of this work can be achieved autonomously (*e.g.* Citeseer, see [3.3.2](#) page [37](#)). The automatic extraction of information from research papers can be helped considerably by having a wide coverage of the material *e.g.* through having a database of names it is easier to pick out the words that may be an author's name.

In summary, the problems this thesis addresses are providing empirical evidence to support open access, by looking for evidence of increased *citation impact*, and the development of tools to aggregate open access material, both for the purposes of collecting data for analysis and to demonstrate the potential of open access as an enabler of new research tools.

## 1.3 Approach

The work reported in this thesis has reflected the needs and the growth of interest in open access within the research and library communities over five years. I started with a relatively small database, with little desire to provide services to the scholarly community. However this has expanded and been driven by a need to support and encourage open access on a global scale – through advocacy, enabling technology and demonstrating the benefit of open access to authors.

My work started in the closing stages of the Open Citation Project ([Hitchcock et al., 2000](#)), a project aimed primarily at providing an add-on service to the arXiv.org for providing hypertext links for references contained in PDF documents (allowing users to click a reference to access the cited work, versus performing a search using the supplied bibliographic data). As well a practical tool for end-users, collecting citation links provides a rich database for analysis of scholarly behaviour. It is this marriage of practical tools and bibliometric analysis that forms the approach taken in this work. Just as having the data prompts analytical questions, so those analyses inform how we might provide new and insightful tools.

I have extensively used data from my Citebase Search service, a collection of some 400,000 predominantly physics papers, their references and usage data. This provides a sizeable, if domain-specific, database to analyse citation behaviour for open access material. Supporting Citebase Search is the Celestial tool, designed to harvest metadata from source repositories using the OAI-PMH. As the focus of the author self-archiving open access strategy has moved from centralised, subject-specific archives towards developing institutionally-hosted archives so it has become increasingly useful to monitor the progress of these disparate resources. The Registry of Open Access Repositories provides a human-compiled listing of archives, and coupled with Celestial, monitors the number of records available from those archives.

In order to test whether open access increases citation impact I have used the ISI Science Citation Index<sup>3</sup> (see 3.2.1, page 32). The ISI SCI is widely used by bibliometricians because the raw database can be licensed for research purposes. Also, as the basis for the *Journal Impact Factor* (see 4.2.1, page 55), the ISI SCI is commonly used for evaluative purposes.

Most of the research I have performed is based on real-time tools I have created to analyse the available databases (Citebase Search, Celestial and the ISI SCI). In order to provide real-time results a lot of data is pre-processed into auxiliary tables *e.g.* to separate open access and non-open access papers in the ISI SCI. A number of web cgi scripts generate graphs or summary tables to provide the actual analyses.

Citebase Search and Celestial have two graphing scripts that provide the mini-graphs used in, respectively, Citebase Search's web interface and the Registry of Open Access Repositories (ROAR). Citebase Search has a main statistical graphing script that provides a range of analyses ranging from the relatively simple (*e.g.* how many papers have been indexed per month) to the complex (*e.g.* the obsolescence of research papers based on citation age). The Correlation Generator is a semi-separate tool that allows me to calculate the correlation between download and citation impact, based on a

---

<sup>3</sup>In collaboration with the Université du Québec à Montréal

number of input parameters. The ISI SCI data is analysed by another graphing tool that can separate the data by ISI subject category.

All of the analytical tools I have created are publicly accessible from the Citebase Search and ROAR sites. Most of these analyses haven't been published, because they're relatively uninteresting or are available elsewhere. However, as they are based on the live databases<sup>4</sup> they continue to be available and up to date.

## 1.4 Terminology

When is a reference a citation? When is an article a paper? The vocabulary used in scholarly communication can vary between the publishing industry, research librarians, information scientists and, not least, research academics. My background in this field is as a researcher and open access advocate, which means the terminology I define in this section may be different to that used in the digital library community.

A **reference** is a bibliographic description of a research work. A reference typically contains the author, year of publication, title, periodical (or conference) title, and pagination (volume, start page, issue *etc.*), *e.g.*:

Hajjem, C., Harnad, S. & Gingras, Y. (2005), Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact', IEEE Data Engineering Bulletin 28(4), 39-47.

When an author cites a work they use a reference to allow the reader to locate the cited work. References (in the sciences) are collected together at the end of a research paper, and are collectively a *bibliography* or *reference section*.

---

<sup>4</sup>Citebase Search, Celestial and ROAR are actively updated. The latest ISI SCI available was to-end 2003.

A **citation** is a reference from one work to another work. For the librarian and academic the citation is typically the location within the text (*e.g.* ‘... in his thesis Brody [4] gave a definition of terms ...’). For the purposes of this thesis I use *citation* to refer to the relationship between two research works (a ‘citation link’), rather than any specific usage by the citing author. It is this relationship that forms the basis for citation analysis.

A **paper** is a ‘primary research article’, insofar that it is an article written by a researcher containing citations to other works. This is compared to **articles** that can encompass a variety of documents that are be found in digital library systems *e.g.* technical reports, letters, theses *etc.* I see the primary goal of open access – and by proxy an open access citation index – to analyse the citation behaviour in research papers.

**Open access** is discussed in some detail in the subsequent chapter, but can be summarised as “free, instant access to the full texts of research papers on the web” [Chan et al. \(2002\)](#).

An **e-print** is a web-accessible version of an author’s research paper. E-prints may be pre- or post- peer-review, uncorrected or not, ‘published’ or not. In general e-prints are papers that are destined for publication in a peer-reviewed journal or conference proceedings. The dictionary definition of **publication** is the ‘act of making public’<sup>5</sup> however, in research, published papers are those that have been peer-reviewed and distributed in a journal or conference proceedings. Reviewing the published papers of an author is the primary means of evaluating a researcher’s performance in many areas of research.

An **e-print archive** (abbreviated to **archive**) is a collection of e-prints made freely available on the web (typically self-deposited by authors). Hence my use of ‘archives’ in this thesis is tightly coupled with the concept of open access. [Crow \(2002\)](#) defined **institutional repositories** to mean “digital collections capturing and preserving the intellectual output of a single or multi-university community”. So an institutional repository (or IR) may contain many different kinds of scholarly material including theses,

---

<sup>5</sup>[dictionary.com](http://dictionary.com)



**Table 1.1:** arXiv sub-archives

| Abbreviation | Name                                     |
|--------------|--|
| astro-ph     | Astrophysics                             |
| cond-mat     | Condensed Matter                         |
| gr-qc        | General Relativity and Quantum Cosmology |
| hep-ex       | Experimental High Energy Physics         |
| hep-lat      | Lattice High Energy Physics              |
| hep-ph       | High Energy Physics Phenomenology        |
| hep-th       | Theoretical High Energy Physics          |
| math-ph      | Mathematical Physics                     |
| nucl-ex      | Nuclear Experiment                       |
| nucl-th      | Nuclear Theory                           |
| physics      | Physics*                                 |
| math         | Mathematics*                             |
| nlin         | Non-linear Sciences                      |
| CoRR         | Computer Science*                        |
| q-bio        | Quantitative Biology*                    |

\*The Computer Science, Mathematics, Physics and Quantitative Biology subjects are each separated into a large number of sub-categories, not listed here.

presentations, teaching materials or data sets. (The OAI-PMH just uses the term **repository** to name the server containing metadata records – an OAI repository can contain anything, not just research material.)

### 1.4.1 arXiv Subject Headings

The arXiv – an archive of author self-archived e-prints – uses a subject hierarchy. These subjects are most commonly referred to using their abbreviations, and are used extensively in this thesis when referring to arXiv subject areas (‘sub-arXivs’) *e.g.* **hep-ph** is *High Energy Physics Phenomenology*.

The list in [Table 1.1](#) is taken from the arXiv home page<sup>6</sup>.

---

<sup>6</sup>arXiv.org <http://arXiv.org/>

## 1.5 Thesis Structure

Chapter 1 of this thesis introduces the motivation for this work, the approach taken and background to the scholarly publishing environment.

Chapter 2 gives background for the scholarly publishing environment, the serials crisis and open access.

Chapter 3 provides a summary of the currently available sources of citation data and how that data is being used, in particular the ACM digital library is an example of citation links being used for navigation, the CiteSeer tool provides a citation index for papers available on the web, the ISI Science Citation Index provides a citation index for the ‘high impact’ scholarly literature and lastly the Open Citation Project (OpCit) is introduced. The work undertaken in OpCit has led to the tools and research presented in this thesis. Lastly the OpenURL/SFX-linking environment, Digital Object Identifiers and the CrossRef consortium are described.

Chapter 4 introduces various statistical techniques for analysing the scholarly literature, *e.g.* the various “rules” that have been defined to describe the distribution of citations. It also describes the Impact Factor, a metric developed by Eugene Garfield and the ISI to evaluate the impact of journals within the scholarly community, based on the number of citations they receive.

Chapter 5 describes the Celestial service. There are now many sources of open access material, from many heterogeneous sources. The Open Archives Initiative Protocol for Metadata Harvesting was developed to allow metadata to be harvested from such disparate sources, using (at least) the Dublin Core set of values – a generalised description of documents. The Celestial service harvests and indexes metadata from OAI-compliant archives. This is a first step to building an open access citation index by allowing the discovery of research papers.

The Registry of Open Access Repositories (ROAR) has been developed as part of this work and forms Chapter 6. The ROAR is a registry of research

archives (collections of research material), categorised by type of archive (*e.g.* institutional), country, along with a record of the number of items in each archive. Using this data the progress of research archives in making research papers open access is evaluated.

Chapter 7 describes the core service – Citebase Search - that underpins the research reported in this thesis. Citebase Search has its origins in the Open Citation Project (Chapter 3), by building a citation index from open access research papers. Citebase Search includes a suite of tools for bibliometric analysis (Chapter 4) that analyse citation patterns, author self-archiving and – experimentally – web usage. Citebase Search is also a heavily used service that provides citation navigation and citation-ranked search results for its source archives (although only a subset of those registered in the ROAR – Chapter 6).

Chapter 8 expands the usage analysis performed by Citebase Search (Chapter 7), in particular whether early-days web usage can be used to predict future citation impact. This is tested by using a correlation generator tool – a web interface that allows the user to adjust the various parameters used to calculate the relationship between the total downloads and total citations to individual papers (*e.g.* to test the predictive power of usage data).

Chapter 9 looks out how open access to research papers (Chapter 2) has affected the behaviour of researchers, based on citing behaviour. In particular how open access increases citation impact (the number of citations to papers, hence the number of citations counted to the papers' authors). Early access to research papers, through authors providing access to pre-prints, has resulted in the citation latency reducing (a measure of the delay in research being published, read, and built upon by others).

The thesis concludes in Chapter 10.

## Chapter 2

# The Scholarly Literature and Open Access

*“If you have an apple and I have an apple and we exchange these apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas.”*

George Bernard Shaw ([Delamothe and Smith, 2001](#))

### 2.1 Background

The scientific journal system serves the need for the dissemination of scientific results and forms the official record of science. The *Philosophical Transactions of the Royal Society* is the longest running scientific journal, first published in 1665<sup>1</sup> (the first scientific journal was the *Journal des Sçavans*). [Harnad et al. \(2004\)](#) estimates there are now 24,000 peer-reviewed journals publishing “about 2.5 million articles per year” (although estimates vary depending on the definition of a peer-reviewed journal). This system has remained much the same since its beginnings, with the migration to electronic (web-based) journals largely resulting in

---

<sup>1</sup>About the Royal Society, <http://www.royalsoc.ac.uk/page.asp?id=2176>

electronic versions of their on-paper counterpart. However, the economic nature of the web – where distribution and copying costs drop to virtually nothing – has encouraged many researchers to re-evaluate how the scholarly communication system should work.

[Ginsparg \(2001\)](#) puts forward the e-print arXiv as showing “some of the possibilities offered by a unified global archive.” He points out that “These e-print archives are entirely scientist-driven and are flexible enough either to co-exist with the pre-existing publication system, or to help it change to something better optimized for each researcher.” Ginsparg lays out a three-layer vision of ‘data, info and knowledge’, where e-print archives form part of a data layer aggregated by information services with overlay ‘knowledge’ services, all visible and usable by the user.

[Harnad \(2001a\)](#) describes a system of institutional e-print archives (so the authors at that institution deposit their works in their institutional archive), harvested by federating services into “global virtual archives”. Underlying such a global virtual archive are archives of the “give-away” research literature. These archives are public, toll-free accessible web repositories of literature deposited by its own authors.

Harnad calls the process by which authors deposit their own works in public web repositories “self-archiving”. Harnad does not refer to archiving in the preservation sense, but simply building a publicly accessible collection of one’s own works. The long-term preservation of digital objects (and especially “born-digital” objects) is the subject of much research but, although related, is not discussed here.

The literature referred to by Harnad as being “give-away” are the reports on research undertaken by its authors before and after peer-review. The reason that this literature is given away, rather than sold, is because; “Researchers publish their findings in order to make an impact on research, not in order to sell their words . . . to make a difference, to build upon the work of others, and to be built upon in turn by others.” So it is in the interest of the author to have the greatest exposure, or impact, for the literature they produce.

What sets apart the literature as referred to by Harnad and, for example, a researcher's web page, is that the research literature is peer-reviewed. Harnad describes peer-review as "...the evaluation and validation of the work of experts by qualified fellow-experts (referees) as a precondition for acceptance and publication". However, this peer-review process does not come for free, even if the author produced the words without expecting direct payment for them. The costs associated with implementing this process (administration, editorial) have traditionally been paid for by publishers who recoup their costs by selling access to the research literature through S/L/P – subscription/licence/pay-per-view. Open access to journals could be achieved by shifting the cost from users (S/L/P) to authors (hence their funders). Indeed, many authors already pay 'page charges' on publication in addition to S/L/P ([Kligfield, 2005](#)). But there is a lot of inertia behind journals and publishing in general, making it difficult for any rapid change to the funding model for the vast majority of journals (if change is possible at all). In the mean-time, self-archiving provides a way to open access without changing the economics.

The progress towards the self-archiving vision has been slow, however. [Kling and McKim \(2000\)](#) describes the growth, or lack of it, in a number of research areas, and suggest reasons for the variability across areas. Kling suggests that the success of arXiv is the result of greater trust by the high-energy and astro- physics areas of pre-peer-review works, in comparison to bio-medical, for example. This trust derives from the technical report culture in physics – the result of large, multi-national projects whose internal communications rely on reports. arXiv is still the largest archive of author self-archived papers and 6 years later the percentage of papers available as open access is estimated by [Hajjem et al. \(2005\)](#) as only 5%-16%.

[Kling and McKim \(2000\)](#) distinguish the growth of electronic (*i.e.* web-based) resources and efforts towards free access. They propose that it is inevitable that scientific research will be conducted and communicated across the web, due to the increasing costs of the current on-paper system, and the attractive pull of new features (web-linking, data stores, and so on).

Indeed, [Cox and Cox \(2003\)](#) found 75% of journals were available online and, in studying the citation patterns of undergraduates between 1996-1999, [Davis and Cohen \(2001\)](#) found a significant shift away from citing books (usually only available on-paper) to purely web resources. It is not certain however that scholarly communication through the web will result in a single, global research archive. [Kling and McKim \(2000\)](#) suggest that the wide variety of services built by differing communities indicates that, if anything, the move to web-based services will be divergent rather than convergent. From the perspective of the building a global virtual research archive it is likely that there will be many community-specific services but with a generic, interoperable service that provides a common entry point to these services, based on the common properties of all research and its literature. While the benefit of web access to research papers has been accepted by the publishing and research communities, the benefit of providing free access hasn't advanced to the same degree.

The first research into whether papers available for free on the web have higher *citation impact* was published in 2000. [Lawrence \(2001\)](#) analysed the difference in citation impact (the number of citations an article receives) between articles freely available on the web and those only available through either toll-access services, or paper-only. Lawrence found that articles available free on the web received on average 2.6 times more citations. Analysing only articles submitted to high-impact conferences (defined as conferences whose articles receive a high number of citations), yielded a difference of 2.9 times.

Lawrence based his statistics on 119,924 Computer Science conference articles indexed by NEC's CiteSeer (a web crawler specialising in indexing, and reference-linking research articles). Lawrence suggests that the greater access afforded to articles by being freely available online is the cause for greater impact of those articles.

The citation advantage (or not) of free access at the journal level was studied by [McVeigh \(2004\)](#). Open access journals (journals that don't charge for access) were most common in the Medical and Life Sciences subjects, with the 'Physics, Engineering & Mathematics' subject rapidly

gaining. McVeigh found OA-based journals were fairing worse than average by journal impact-factor, but to a lesser degree in the ‘immediacy index’ (the speed with which those citations are made). While McVeigh doesn’t provide any head-line results he did find four relatively new journals (launched in the ‘last 10 years’) were ranked in the top 10% – which represents quite an achievement for a new journal.

Free access should lead to more citations: the more chance someone has of finding an author’s work, the more likely it is they will cite it. Conversely if there is no access to a work it can never be cited. If all research papers were freely accessible to all would-be users it is uncertain whether the relative citation impact between papers would change. It is possible that citations will be focused on higher quality articles versus those that are more accessible – that a paper is published in a widely-subscribed to serial will be less of a factor than whether that paper is well received by the community of interest. Hence the focus of research into a relative citation advantage for open access papers focuses on those fields that are in a *transitional* stage, with a sufficient body of open access literature to test for a significant difference but not yet 100% open access.

## 2.2 Scholarly Publishing in Transition

While the printing press allowed authors to distribute their ideas far wider than hand-copying (or verbal tradition) allowed, the web has fundamentally changed the economics of information. Where once, if you wanted access to information you had to pay a toll-charge in the form of subscriptions (to cover the cost of distributing a physical, printed object) the web allows any number of users to obtain a copy for little more than the initial cost of creating that information. Email and personal web pages allow scholars to communicate and collaborate directly, without the delays and cost associated with printing and posting.

Despite the cheap distribution nature of the web, research publication remains focused on the mechanics and economic model of paper-based



publications. Unlike most creative authors, research authors are not paid to write research papers, instead they publish their work in order for others to read and build upon what they have done. The reward for research authors is through the influence and impact they gain – in effect the publications act as an advertisement for the authors. And yet most of the current research literature published through journals charges the user – through subscriptions, licences and pay-per-view (S/L/P), which is at odds with the interest of the author. In effect S/L/P's primary purpose is to *deny* access to only those who can pay, while the author's interest lies in *allowing* access to as many users as possible. In the on-paper world it makes sense to charge the user, to cover the cost of printing and distribution, but on the web an up-front fee could cover the cost of the creating the paper with access provided for free to all would-be users.

Authors are beginning to recognise the potential to bring the distribution power of the web to improve access to their work. arXiv (Ginsparg, 1996, 2001) is the largest author self-archived archive of research papers. Started in 1991, arXiv has grown to become one of the most important resources for experimental physicists (now with 54,000 submissions annually<sup>2</sup>). Swan and Brown (2005) found “almost half (49%) of [authors that responded] have self-archived at least one article during the last three years.” Schroter (2006) found three quarters of *BMJ* authors surveyed thought free access to their papers “very important or important” to their decision to submit to the *BMJ* (which provides open access to research papers, but not other content). Despite this Hajjem et al. (2005) found only 5%-16% of papers were available on the web as open access.

While many authors are aware of the improved visibility and impact that open access can provide to their work, within the libraries community the spiralling cost of providing access to research journals has pushed the open access agenda.

---

<sup>2</sup>Monthly submissions to arXiv [http://arxiv.org/show\\_monthly\\_submissions](http://arxiv.org/show_monthly_submissions)

## 2.3 The ‘Serials Crisis’

[Odlyzko \(1995\)](#) states that “in 1870 there were only about 840 papers published in mathematics. Today, about fifty thousand papers are published annually.” The rate of increase in the number, size and cost of journals has outstripped the increase in library budgets, leading libraries to sacrifice the purchase of serials and monographs and to increasingly ration subscriptions to only those journals with the highest usage (or at least perceived usage). The *serials crisis* has a number of causes, the biggest of which is the increase in material published. However, above-inflation increases in journal prices (particularly by commercial publishers – [White and Creaser \(2004\)](#)) and decreasing library budgets have also had an effect.

The transitional process from paper-publications to electronic has resulted in several commercial publishers offering bundling of journals – the so-called ‘Big Deal’. Bundles of journals (*e.g.* within a subject) are sold as a block unit. This makes commercial sense for publishers selling electronic versions, where the incremental cost to increasing access is marginal and has been attractive to some libraries, as they can purchase more titles with the same budget. [Branin et al. \(1999\)](#) argues strongly against accepting the big deal as it reduces libraries’ ability to choose the journals they actually *need* – “In the longer run, these contracts will weaken the power of librarians and consumers to influence scholarly communication systems in the future. Librarians will lose the opportunity to shape the content or quality of journal literature through the selection process.”

In the face of rising journal costs some institutions have looked to pressure commercial publishers to reduce their charges, *e.g.* at Stanford University, where the cost of journal subscriptions had risen 50% in 5 years ([Miller, 2004](#)), the library was encouraged to “systematically drop journals that are unconscionably or disproportionately expensive or inflationary. Special attention should be paid to Elsevier.”

The reality is that the journal subscription charges made by publishers to libraries, while the most easily identifiable cost in the system, is only a small

part of the total cost of publishing an article. [Odlyzko \(1998\)](#) estimated that publishers' revenue accounts for \$4,000 on average per paper, compared to \$8,000 for "library costs other than purchase of journals and books", \$4,000 for "editorial and refereeing costs" and \$20,000 "authors' costs of preparing a paper". While the costs of authoring, editing and reviewing research papers have scaled with the increasing amount of investment in research, the budget for libraries has remained relatively static. So, because the publishing industry is only a small part of the cost of publishing a paper, reducing the cost of journal subscriptions will only help in the short term.

## 2.4 Open Access to Research Papers

Open access can help the serials crisis both in the short and long term. Author self-archiving provides a parallel system to access research papers (for those without a subscription), that resolves the immediate *access* problem. In the longer term changing the economic model of research publication towards a front-loaded ('author-pays' model) will allow the cost of publication to scale with the level of investment in research.

### 2.4.1 Author Self-Archiving

Author self-archiving is a parallel process to publishing in a journal. By providing an alternative – free – means of access to the author's paper authors maximise the number of potential users of their work. [Pinfield \(2004\)](#) stated that author self-archiving "has the potential to revolutionise scholarly communication, making it more efficient and effective." It is in the author's interest to have their work widely distributed, read and used by other researchers – in essence maximising research impact through maximising access. While research authors have for a long time out-sourced the distribution of their work to publishers, in return for granting an exclusive right to distribute their work, the web allows authors to perform this role themselves.

Since the advent of the web authors have been uploading versions of their work onto the web, distributing by email, or otherwise giving access to their work to colleagues through the internet. Author self-archiving can be achieved in three ways: author self-archival in a centralised (typically subject-specific) repository, in an institutional repository or on a personal web site. The version of the paper uploaded by an author is usually a pre-published version *i.e.* prior to the formatting by the publisher according to the journal's style. This version is likely to be the most readily available version (but still includes peer-reviewer's corrections) but is also more acceptable to the publisher, as it isn't the 'official' journal version.

Should a journal's policy prohibit the author from self-archiving the post-print version (*i.e.* following peer-review) ([Harnad, 2001b](#), section 6) proposed the "pre-print/corrigenda strategy". Briefly, this strategy is to self-archive a pre-print (the version before peer-reviewing) and subsequently to append a list of corrigenda. Thus the user is made aware of any errors caught by the peer-review process, but without having to use the publisher's version.

### 2.4.2 Institutional Repositories

Is the purpose of open access only to provide an alternative, free access to a version of the journal literature (hence served as well by a web site, as a complex digital library system), or are institutional repositories an end of themselves? As research based on open access literature, my work would not be possible without open access to the scholarly literature, which is achievable through a simple web site. But my work would be made much more difficult were it not for the OAI-PMH interface and consistency afforded by institutional repositories. On the other hand were these repositories to be more complex – *e.g.* by demanding structured citation data from authors – so building an open access citation index would be made that much easier. But how many authors would be prepared to spend the time to mark-up their work, on what is essentially a side-line activity to journal publication? For now, no institutional repositories are capturing

structured citation data from authors<sup>3</sup>, so citation linking is dependent on services creating the structured data from the authors' full-text.

[Lynch \(2003\)](#) defines an institutional repository as “a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members.” He goes on to say institutional repositories “will contain the intellectual works of faculty and students—both research and teaching materials—and also documentation of the activities of the institution itself...” Lynch’s vision of institutional repositories is very broad and doesn’t address perhaps the greatest risk in implementing institutional repositories: the institutional repository doesn’t attract any content. For example, [Foster and Gibbons \(2005\)](#) cite MIT Libraries as estimating an annual maintenance cost for their institutional repository of \$285,000, but with “approximately 4,000 items currently in their IR, that is over \$71 spent per item, per year.” I investigate the uptake of institutional repositories in [chapter 6](#), page 71.

The cost of establishing an institutional repository varies greatly across implementations. [Morrison \(2005\)](#) points out that “At the low end of the cost range is the completely free institutional repository [software] ... At the higher end of the cost range, a large university could plan a comprehensive institutional repository program, not only for the open access research literature, but also for all manner of other types of information.” The irony being for larger institutions that they can end up pricing themselves out of building a repository as the perceived requirements spiral out of control. The free end of repositories is likely to be highly focused on providing an efficient infrastructure for authors to post their work (perhaps with some minimal checking of validity/relevance) – [Carr and Harnad \(2005\)](#) found that authors were taking about 10 minutes to deposited a paper (in a Computer Science department). The high-end repository might demand far more detailed information from the author to support the generation of detailed reports on research output *e.g.* in the U.K. to support

---

<sup>3</sup>By default the GNU EPrints software asks the user to copy-and-paste their references into a text field, but doesn’t convert those references into structured data usable for citation linking.

the Research Assessment Exercise ([Carr and MacColl, 2005](#); [Day, 2004](#)).

### 2.4.3 Open Access Journals

The economics of the web makes distribution virtually cost-free.

Distributing data on the web is very cheap once one person has paid for and decides to give their copy away *e.g.* by putting the copy on a web site. This has caused problems for the music and movie industries, but the difference between research papers and other intellectual media is that the authors of research papers don't expect royalties. As long as research authors' work is attributed and not modified it is in their interest to be as widely copied as possible. Anyone can post any work they like onto the web, to be accessible by anyone with web access, what makes the difference with journals is that the work is *peer-reviewed*. Peer-review is a check and filter for research, both to catch bad research and a stamp of quality (depending on the prestige of the journal). Regardless of how the costs of editing, administration and technical production of a journal are covered, peer-review is the difference between scholarly communication and general information available from the web.

The primary source of revenue for subscription-based journals is charging users for access. [Kaufman-Wills Group \(2005\)](#) looked at how a range of open access journals were funded. In the preface they state “few of the Open Access journals raise any author-side charges at all; in fact, author charges are considerably more common (in the form of page charges, colour charges, reprint charges, *etc.*) among subscription journals.”

Of the two largest commercial open access publishers in the survey (BMC<sup>4</sup> and ISP<sup>5</sup>), BMC charged a publication charge to authors and ISP covered it's costs from industry grants. While most of the journals studied in [Kaufman-Wills Group \(2005\)](#) did not charge up-front fees for publication it is difficult to imagine ad-hoc grants and sponsorship could scale to the rest

---

<sup>4</sup>Biomed Central <http://www.biomedcentral.com>

<sup>5</sup>Internet Scientific Publications <http://www.ispub.com/>

of the publishing industry. BMC and PLOS<sup>6</sup> have set up new publishing businesses where authors are charged a publication charge on acceptance for publication *i.e.* post peer-review. (Few authors would personally pay the charge, instead their institution would pay either per-paper or by an institutional membership.) Springer<sup>7</sup> have created an ‘open choice’ system where authors can opt to pay for open access, thus mixing open access and subscription content within individual journals.

*Nucleic Acids Research* (NAR) is a journal published by Oxford University Press that changed from a subscription-based to author-charge (open access) revenue model. Richardson (2006) reports how changing to an immediate open access model (in January 2006) has affected NAR. In a user survey it was found 88% of respondents “agreed or strongly agreed that the principle of free access for all is important.” In assessing the impact of open access on usage patterns evidence was found “of the phenomenal impact of search engines [on] NAR usage”. While the back-catalog of NAR content was made freely available from 2003 (but with a 6-month embargo on new papers) immediate open access “simply added a little to the massive growth that was already going on” - by an estimated 7-8%.

The Directory of Open Access Journals<sup>8</sup> is a registry of a little over 2,000 (as of January 2006) “free, full text, quality controlled scientific and scholarly journals”. Harnad et al. (2004) estimates there are 24,000 peer-reviewed journals, which means open access journals account for only a small proportion of the total.

## 2.5 Policy Makers and Open Access

From October 2005 the Wellcome Trust – the “UK’s largest non-governmental source of funds for biomedical research<sup>9</sup>” – required all grantees to deposit any research papers resulting from Wellcome Trust funded research to be

---

<sup>6</sup>Public Library of Science <http://www.plos.org/>

<sup>7</sup>Springer <http://www.springer.com/>

<sup>8</sup>Directory of Open Access Journals <http://www.doaj.org/>

<sup>9</sup>Wellcome Trust – About Us <http://www.wellcome.ac.uk/aboutus/>

deposited (hence made publicly accessible) into the PubMed Central service<sup>10</sup> (Wellcome Trust, 2006). In addition the Wellcome Trust has undertaken to cover the author charges made by open access publishers.

Reports by the UK Parliament Science and Technology Committee Publications (2004) and the US House Appropriations Committee (2004) have recommended mandating that researchers provide open access to their research articles by self-archiving them free for all on the web. The UK committee acts in an advisory role to the UK government on science and technology matters and can only provide recommendations. The US committee has much greater influence as its recommendations go into how the budget for the large US medical funding councils is set (*e.g.* the National Institutes of Health<sup>11</sup>).

The UK report recommended that “all UK higher education institutions establish institutional repositories on which their published output can be stored and from which it can be read, free of charge, online.” It further recommended that experiments be undertaken in alternative open-access based funding models for journal publication, perhaps using an author-pays model (where the author pays the publishing cost, with the result that the paper can be given away for-free on the web). While these recommendations haven’t been pursued by the UK government, in its response it didn’t preclude the agencies it funds – *e.g.* Research Councils UK (RCUK) – from mandating open access (Prosser, 2004).

Government policy on open access in the US has been focused on medical research, with the US House Report resulting in an NIH policy of requesting all NIH-funded researchers place a copy of their research papers into a central repository within twelve months of publication (Suber, 2006). But this voluntary policy has resulted in less than 4% of the potential papers being made open access (Tanne, 2006).

While both the UK and US governments have recognised the benefit of increased access to research papers, in particular that tax-payers shouldn’t

---

<sup>10</sup>PubMed Central <http://www.pubmedcentral.nih.gov/>

<sup>11</sup>National Institutes of Health <http://www.nih.gov/>



have to pay to access the results of research that they fund in the first place, they are yet to mandate open access to the research they fund. However, that open access has been brought up at the government policy level at all demonstrates the importance that increasing access to research has achieved. The Wellcome Trust policy demonstrates they see the benefits of open access outweighing the (potential) financial costs to funding agencies. It is therefore easy to see government's encouraging their tax-payer funded research agencies to adopt policies mandating open access.

## 2.6 Enabling Open Access using the Web

Although the Serials Crisis helped draw attention to open access, a deeper problem has been identified: the research access/impact crisis. No institution has the funds to subscribe to every journal that is published. If the University of Harvard – with 100,000 serials – is the benchmark for how many serials are available, researchers at other institutions are missing access to a lot of material (see [Kyrillidou and Young, 2004](#), table 3) *e.g.* rank-ordered the 20th University, Johns Hopkins, has 50,000. Even if every journal were sold at-cost most libraries couldn't afford most of the available journals. Yet every potential user that an article loses is lost potential impact for it's author, the author's institution, the research-funder and for research itself. This lost impact is the access/impact problem, and the advent of the web itself has provided the solution by enabling open access [Harnad et al. \(2004\)](#).

The web has revolutionised the dissemination of information but only half of UK authors (see [Swan and Brown, 2005](#)) have used the web's power to maximise the visibility, accessibility and usage of their work by providing open access through self-archiving. [Lawrence \(2001\)](#) found that papers in Computer Science that were freely accessible on the web received 2.6-2.9 times as many citations as their subscription-only counterparts. I have performed a similar analysis for physics and mathematics papers self-archived in arXiv and have found that open access papers (on average)

are cited 2-3 times more than papers available only a subscription-based journal (Harnad and Brody, 2004) – this analysis is expanded in [section 9.2](#).

My findings are based on comparing open access papers with non-open access papers within the same subject and, where there is sufficient data to test, the same journal and year. The greater citation impact of open access papers is a *competitive* advantage – should any field reach 100% open access there would not longer be any non-open access papers to have an advantage against. Kurtz<sup>12</sup> has shown that in astrophysics – a field in which there is already effectively 100% open access through institutional licensing – overall usage of papers is doubled over what it was before open access. The increase in *usage* in a field where 100% of authors already have access suggests that there are wider benefits for open access than just research-authors.

There are other benefits to open access than just making access cheaper for users. Subscription-based access requires users to access papers through either the publisher's web site or an aggregating service that licences the content. Although Google has managed to index a number of publishers' content in their free Google Scholar service (Google, 2005; Sullivan, 2004). Because subscription-based access depends on *denying* access so the number and quality of services available to users is also limited.

Open access allows any individual or organisation to harvest and index open access papers, so the monopoly on full-text access is broken. Even with the small percentage of open access available today, there are already at least 10 different service providers using the OAI-PMH standard, which makes all OAI compliant Archives (whether they consist of self-archived institutional output or journal/publisher databases) interoperable with one another (Van de Sompel and Lagoze, 2000). The major web search companies have shown interest in building scholarly search tools (*e.g.* Google Scholar<sup>13</sup>). I expect a range of aggregating services will be developed – from toll-free, generic web search engines such as Yahoo!<sup>14</sup>, to specialised toll-based services such as

---

<sup>12</sup>See slide 25 in “Self-archiving Illustration”, presentation by Stevan Harnad <http://www.ecs.soton.ac.uk/~harnad/Temp/daser-harnad.ppt>

<sup>13</sup>Google Scholar <http://scholar.google.com/>

<sup>14</sup>Yahoo! search <http://www.yahoo.com/>

Elsevier's Scopus<sup>15</sup> which provides categorised navigation and augmented metadata. In addition to meta-searches, open access opens the possibility of designing services that analyse patterns in scholarly research, using the built-in citation links, without being limited to proprietary databases that cover only a portion of the total literature.

Open access is an exciting development for bibliometricians (researchers that study libraries' data). The most comprehensive citation database available today (by total citations) is the ISI Science Citation Index (see [subsection 3.2.1](#), page 32). It covers around 8,700 ([ISI, 2004](#)) of the world's 24,000 total journals ([Harnad et al., 2004](#)) but is closed and provides limited analytical tools. With autonomous open access citation tools (*e.g.* CiteSeer [Lawrence et al., 1999](#)) information scientists can now build comprehensive citation databases limited only by what has been made open access to date.

Citation databases allow the literature to be navigated backwards and forwards in time by following citations to and from any paper or – using co-citation analysis – to find related papers (which papers cite the same papers? Or are cited by the same papers?). Citation analysis can be used to find emerging fields, to map the time-course and direction of research progress and to identify synergies between different disciplines. Content analysis of the full-text content of papers can be used in similar ways to deepen the analysis of the underlying patterns, as well as to aid navigation, search and evaluation.

While many publishers currently provide web links for citations (*e.g.* see [section 3.2](#), page 31), these links are often dependent on bespoke implementations for particular cited publisher's services or the user may not have a subscription to the cited paper. Open access will enable a service to gather and link to any open access material, potentially allowing all cited papers to be linked to (and be immediately accessible to the user following the link).

---

<sup>15</sup>Scopus Info <http://www.info.scopus.com/>

## 2.7 Evaluating Research Performance in Open Access

How research is evaluated is a sensitive subject, especially for the researchers being evaluated, but “Research evaluation has emerged as a key issue in many industrialized countries, where universities are faced with demands for greater accountability and the consequences of diminished funding. Universities today are expected to be both efficient and accountable. These pressures have made evaluation essential.” (Geuna and Martin, 2003) For researchers peer-review is perhaps the most accepted form of evaluation – as Day (2004) points out “One of the advantages of peer review over other approaches is its widespread use elsewhere in the academic world, *e.g.* as part of the publication process and for deciding the allocation of research grants.” The problem of using peer-review for evaluation is it is both *costly* and a *subjective* measure of quality.

Quantitative measures (*e.g.* citation impact) can be calculated cheaply, given a citation database. Oppenheim (1996) performed a comparative study of the results from the 1992 U.K. Research Assessment Exercise (RAE) and the ranking resulting from counting citations. He found a correlation of  $r = 0.82$  between the RAE ranking and average citation counts per member of staff for research departments. Oppenheim estimated his approach would cost “about one thousandth of the cost of conducting the RAE.” Despite this the 2008 RAE will follow the same costly, peer-review based process (UK RAE, 2006).

Open access could provide three benefits for research evaluation. Firstly, institutional repositories could provide a platform for the administration of collecting and submitting for evaluation research papers and other ‘indicators of prestige’ (Carr and MacColl, 2005). (To date the effort that has gone into making RAE submissions is largely used for only one purpose – the national evaluation of research departments. By capturing the RAE data in an institutional repository that data could be reused for many purposes *e.g.* for author publication records or for advertising an

institution's research output.)

Secondly evaluation could be an opportunity to capture *all*<sup>16</sup> the research publications of authors and – by making them open access – evaluation tools could harvest and evaluate all research output automatically and autonomously. Not only would this provide a more *comprehensive* evaluation but, by making all the research open access, would provide all the benefit of open access to papers.

Thirdly by making all of the information used for evaluation (the full-text papers and prestige measures) open access anyone will be able to evaluate the method of evaluation and to propose or build new methods. Currently the RAE is not *empirically* tested – by making all of the data available it may be possible to test how effective the RAE is at improving the quality and efficiency of research (*e.g.* by performing a time-series analysis of citation impact).

## 2.8 Conclusion

The online era has not produced a substitute for the traditional research publication system, but a powerful new supplement to it, particularly in the area of access provision and research evaluation, based on impact. What is needed now is for institutions and research funders to provide the tools (institutional repositories and open access journals), encouragement and, if necessary, mandates to authors to provide open access to their research publications.

---

<sup>16</sup>The RAE looks at only 4 papers per researcher – ([UK RAE, 2006](#))

## Chapter 3

# Citation Data and Analysis Tools

### 3.1 Introduction

The majority of the data presented in this thesis is derived from full-text papers posted by their authors in publicly accessible repositories. This author ‘self-archived’ literature is typically research papers destined for publication in peer-reviewed journals or conference proceedings, but aren’t the publisher’s final (edited) version. This literature is therefore a mix of pre- and post- peer-review papers and technical reports. Despite this ambiguity the largest collection of author self-archived literature – the physics, maths and computer science arXiv – is heavily used, and authors write papers explicitly citing pre-published e-prints (identifiable because they only use the arXiv identifier).

### 3.2 Commercial Services

Most commercial digital library services now provide some level of citation linking support. These citation links typically provide internal citation links

(links between papers in the same collection), with some links to external collections *e.g.* using DOIs (see [subsection 3.5.3](#)) or to subject-specific tools such as PubMed.

The difficulty with commercial digital library services is they do not provide easy access to the underlying citation link databases. While the ISI WoS will license their data for bibliometric research, no similar licensing is available from other providers. While CrossRef is the most widely used linking service by publishers, they don't provide their database for use by bibliometricians. Indeed CrossRef is normally only queried for the DOI to link against, rather than the publisher downloading the database.

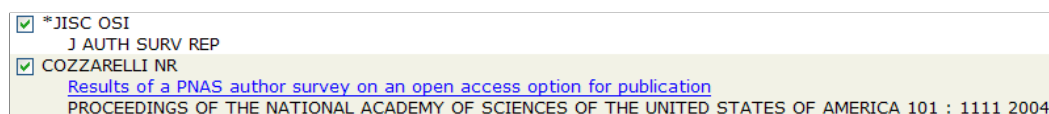
The lack of available citation data is a hindrance for bibliometric research, in particular because it makes comparing the quality and coverage of citation link databases difficult – the only studies possible have to be performed by hand, hence limit the sample size (*e.g.* a single journal in [Bauer and Bakkalbasi, 2005](#)).

### 3.2.1 The ISI Web of Science

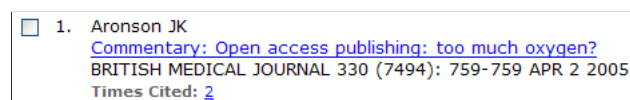
The ISI Web of Science is the combined, web version of ISI's *Science Citation Index*, *Social Sciences Citation Index* and *Arts & Humanities Citation Index* databases. The Science Citation Index (SCI) was first published in 1964 ([Yancey, 2005](#)). The idea for the Science Citation Index has its origins some nine years earlier – [Garfield \(1955\)](#) proposed a “a bibliographic system for science literature that can eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making it possible for the conscientious scholar to be aware of criticisms of earlier papers.”

Garfield realised his proposal by setting up the *Institute for Scientific Information* who built and published the Science Citation Index (described in [Garfield, 1964](#)), of which the first version indexed 613 journals and 1.4 million citations ([Yancey, 2005](#)). The Social Sciences (1970-) and Arts & Humanities (1975-) indexes expand ISI's coverage of citation data to those respective subjects.

A citation index works by allowing the user to easily locate all the papers that cite a given paper. A citation is a one-way link (because it is only in the citing paper) but, by indexing all papers, a citation index can locate all citing papers. For example using the Web of Science I can search for a paper “Perceptions of open access publishing: interviews with journal authors” by *Shroter, S* and find papers that cite that paper or follow links to cited papers. In effect this allows the user to navigate forwards and backwards in time by following citations (as cited papers are older and citing papers are newer). **Figure 3.2** shows two example entries from the bibliography. The Web of Science shows all references in their abbreviated form and, where it also has a record for the cited item, provides a link to the cited paper (including the title for the cited paper, that wasn’t in the original reference). **Figure 3.1** shows an example from the ‘Cited References’ (papers citing the current paper), that provides a link to the citing paper along with its bibliographic information.



**Figure 3.1:** ‘Cited References’ in the ISI Web of Science



**Figure 3.2:** ‘Citing Articles’ in the ISI Web of Science

The Web of Science doesn’t publish research papers itself but provides an aggregated index of many other publishers. [Atkins \(1999\)](#) describes how the Web of Science builds a citation index from the information supplied by publishers. The Web of Science combines human and autonomous systems to input certain key fields from a reference (*e.g.* the page a paper appears on in the journal) and compares these keys to a database of existing papers. As references are input they are looked up, if a match results it is most likely the reference is correct or, if no match results, a later process attempts to resolve any mistakes by the author, or to be certain that the reference does not exist in the Web of Science.



### 3.2.2 The ACM Digital Library

The ACM Digital Library<sup>1</sup> includes “full text to ACM publications, over 10-years worth of tables of contents for its journals and conference proceedings, bibliographic reference pages, and free-text search facilities for bibliographic material ... and links to full text where available” (Denning, 1997). The library was later enhanced so that the “reference section of an article will either link to other articles within the Digital Library or to appropriate sites outside the ACM Digital Library, if the reference corresponds to non-ACM material” (White, 1999).

As well as linking references the Library lists ACM papers citing the current paper (‘citation analysis’). The ACM has used OCR (Optical Character Recognition) to scan the references from existing papers in its digital library (Bergmark et al., 2001; White, 2001), and then linked those references to papers found in the Library. Figure 3.3 shows these links for a paper in the Library, with links on references that were found, and a ‘Citings’ section that shows papers found that cite the current paper. Similarly to the Web of Science all references are shown, but using the original author’s text rather than an abbreviated/normalised form, with references linked where the cited paper is also in the Library.

The citation database behind the ACM Digital Library isn’t open access. Although the web site is indexed by Google (hence following a search result from Google to the ACM is free), following links within the ACM requires a subscription. Access to the ACM Digital Library is possible for research purposes but I haven’t pursued that in this thesis.

### 3.2.3 Elsevier’s ScienceDirect

ScienceDirect<sup>2</sup> is the content delivery platform for Elsevier Science (*i.e.* access to journals and other material published by Elsevier). ScienceDirect

---

<sup>1</sup>ACM Digital Library <http://www.acm.org/dl/>

<sup>2</sup>ScienceDirect <http://www.sciencedirect.com/>

↑ **REFERENCES**

Note: OCR errors may be found in this Reference List extracted from the full text article. ACM has opted to expose the complete List rather than only correct and linked references.

- 1 Brody, T and Hickman I. (2000) Mining the Social Life of an Archive. OpCit Internal Technical Report. <http://opcit.eprints.org/tdb198/opcit/>
- 2 Carr, L. (2001) The Use of Open Archives: Who, How Often and Why. Presentation at Open Archives Workshop, European Conference on Digital Libraries 2001. <http://www.ecs.soton.ac.uk/~lac/opcit.who>
- 3 Rune Dalgaard, [Hypertext and the scholarly archive: intertexts, paratexts and metatexts at work, Proceedings of the twelfth ACM conference on Hypertext and Hypermedia, August 14-18, 2001, Århus, none, Denmark](#)
- 4 Carl Lagoze, Herbert Van de Sompel, [The open archives initiative: building a low-barrier interoperability framework, Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, p.54-62, January 2001, Roanoke, Virginia, United States](#)

↑ **CITINGS 2**

[Leslie Carr, Timothy Miles-Board, Gary Wills, Guillermo Power, Christopher Bailey, Wendy Hall, Simon Grange, Extending the role of the digital library: computer support for creating articles, Proceedings of the fifteenth ACM conference on Hypertext & hypermedia, August 09-13, 2004, Santa Cruz, CA, USA](#)

[Mike Thelwall, Gareth Harries, Do the Web sites of higher rated scholars have significantly more online impact?, Journal of the American Society for Information Science and Technology, v.55 n.2, p.149-159, January 15, 2004](#)

**Figure 3.3:** Citation links in the ACM Digital Library.

claims<sup>3</sup> to have “over 25% of the world’s science, technology and medicine full text and bibliographic information” – corresponding to some “6.75 million articles<sup>4</sup>.”

Figure 3.4 shows an extract from the bibliography for a paper in ScienceDirect. Each reference is shown as written by the author, but with multiple links (where possible) to the cited paper. These include the ‘Full Text + Links’ or PDF if it’s in ScienceDirect, links to Elsevier’s Scopus citation index, links to the MEDLINE database or to retrieve the cited paper using CrossRef (if the cited paper has a DOI). 3.5 shows an extract for citing papers in ScienceDirect, which is obviously restricted to only those journals published on the ScienceDirect platform.

### 3.3 Open Access to Citation Data

While there are many commercial (subscription-based) services now offering citation linking, several open access tools have also appeared, some of which

<sup>3</sup>Content on ScienceDirect <http://info.sciencedirect.com/content/>

<sup>4</sup>On March 2006 ScienceDirect contained 7.46 million full-texts, according to the home page.

[12](#) P.S. Tamber, F. Godlee and P. Newmark, Open access to peer-reviewed research: making it happen, *Lancet* **362** (2003), pp. 1575–1577. [SummaryPlus](#) | [Full Text + Links](#) | [PDF \(68 K\)](#) | [Abstract + References in Scopus](#) | [Cited By in Scopus](#)

[13](#) Gibson B. From Transfer to Transformation: Rethinking the Relationship between Research and Policy. PhD thesis, Australian National University, 2003.

[14](#) S. Schroter, L. Tite and R. Smith, Perceptions of open access publishing: interviews with journal authors, *BMJ* **330** (2005), p. 756. [Abstract-MEDLINE](#) | [Abstract-Elsevier BIOBASE](#) | [Abstract-EMBASE](#) | [Order Document](#) | [Full Text via CrossRef](#) | [Abstract + References in Scopus](#) | [Cited By in Scopus](#)

**Figure 3.4:** References are linked where the cited item is known to ScienceDirect.

3. ☐ **Can we achieve health information for all by 2015?** • ARTICLE  
*The Lancet*, Volume 364, Issue 9430, 17 July 2004-23 July 2004, Pages 295-300  
 Fiona Godlee, Neil Pakenham-Walsh, Dan Ncayiyana, Barbara Cohen and Abel Packer  
[SummaryPlus](#) | [Full Text + Links](#) | [PDF \(151 K\)](#)

4. ☐ **Open-access publishing** • CORRESPONDENCE  
*The Lancet*, Volume 364, Issue 9428, 3 July 2004-9 July 2004, Pages 25-26  
 Anthony Costello and David Osrin  
[SummaryPlus](#) | [Full Text + Links](#) | [PDF \(67 K\)](#)

**Figure 3.5:** Following the ‘Cited By’ link shows a list of papers citing the current paper.

provide access to the underlying citation database. Citeseer is the most widely recognised open access citation index – certainly in Computer Science – however there are now several other open access projects and tools for citation indexing. This section expands on three developments: RePEc, Citeseer and the Open Citation Project.

### 3.3.1 RePEc

The RePEc/WoPEc<sup>5</sup> service is an index and archive of economics research papers, started in 1993 (Karlsson and Krichel, 1999). RePEc is a collection of repositories that together form a virtual collection on which services have been built. Two notable services are LogEc and CitEc<sup>6</sup>, that respectively index accesses and citations to RePEc papers.

CiTeC uses CiteSeer algorithms to process, parse and link citations from papers in RePEc. CiTeC has processed 74,979 papers, resulting in 1,667,669

<sup>5</sup>RePEc: Research Papers in Economics <http://repec.org/>

<sup>6</sup>CitEc: Citations in Economics <http://citec.repec.org/>

references of which 535,080 have been linked to the cited paper. CiTeC doesn't provide any end-user services itself but instead provides that data to other services within the RePEc family *e.g.* the IDEAS<sup>7</sup> service that allows authors to find out how many citations they and their papers have received. The citation data in CiTeC isn't currently exported by the RePEc OAI-PMH interface (see [section 3.4](#)).

### 3.3.2 Citeseer

[Lawrence et al. \(1999\)](#) describe how Citeseer autonomously (*i.e.* without human intervention) crawls the web for research literature, defined as papers that contain a bibliography, parse out the references from the full-text, and then perform reference linking against the existing Citeseer database.

Citeseer uses existing web search engines to locate possible papers. It retrieves these papers, converting them to plain text ready to be parsed. Citeseer locates the reference section, and then parses the references using an invariants-first heuristic method. This identifies common aspects of all the references (*e.g.* a reference number), and then extracts each invariant field from each of the references *e.g.* a year of publication. Each of these reference fields can then be used as a fuzzy query over the existing Citeseer database (*e.g.* author and title, or year and title). The fuzzy query combines fields in order to build clusters of similar references and paper citations (the reference metadata of an actual paper). This allows Citeseer to work-around errors in authored references, or different references to the same paper if the paper was published in more than one location.

In addition to providing reference links, which allows the user to easily navigate between papers, Citeseer provides citation analysis (links to papers that cite the currently viewed paper), as well as co-cited papers (links to papers that have been cited alongside the current paper). Citeseer provides the user with the context that an paper was cited in - the text in the body of the paper around the citation, *e.g.* 'In [30] the authors describe African

---

<sup>7</sup>IDEAS: Economics and Finance Research <http://ideas.repec.org/>

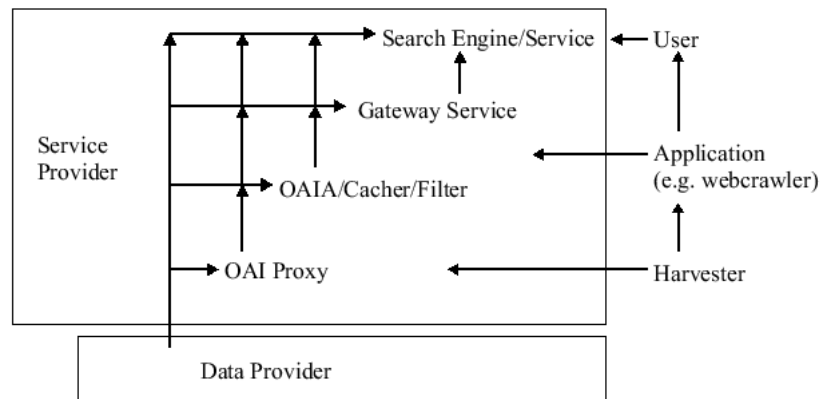
economics’.

### 3.3.3 The Open Citation Project

[Hitchcock et al. \(2000\)](#) describes the Open Citation Project (OpCit), which provides reference linking over existing e-print repositories. [Hitchcock et al. \(2002\)](#) describe how Citebase Search (see [chapter 7](#)) – developed as part of OpCit – interacts with a source archive, providing reference linking services to end users, as well as exposing those references links back to the source archive. This exposure is through an interface using the OAI-PMH, so any third party service could harvest this data to build extended services.

[Liu et al. \(2002\)](#) propose extending these two parties to a larger infrastructure of interacting tools, each providing services over the same papers, but all interoperable and navigable by the user. Source repositories are harvested by intermediate caching, federating, and gateway services. These process the source data, augmenting it (*e.g.* adding linked references), or normalising (*e.g.* converting dates to a standard format). This augmented, normalised data can then be used by end-user services which can provide multiple services for the same records. Using unique identifiers these multiple services can be linked together to provide a richer, virtual service to the user.

In the [Liu et al. \(2002\)](#) infrastructure an OpCit service is a ‘gateway service’ (see [Figure 3.6](#)) that harvests data from repositories (or other OAI-PMH-compatible services), along with full texts, and then re-exposes this data through its own OAI-PMH interface for other services to process, or provide to end-users. This framework is a similar vision to that implemented in RePEc/CiTeC, but is yet to be more widely used. Partly this is due to the lack of standards for the transfer of bibliographic data (although OpenURL is now available, see [subsection 3.5.2](#)) but also to a – quite correct – focus on getting content into source repositories. However, now that source repositories are getting established (see [chapter 6](#)) interest is growing in building more complex end-user services, enabled through the OAI-PMH.



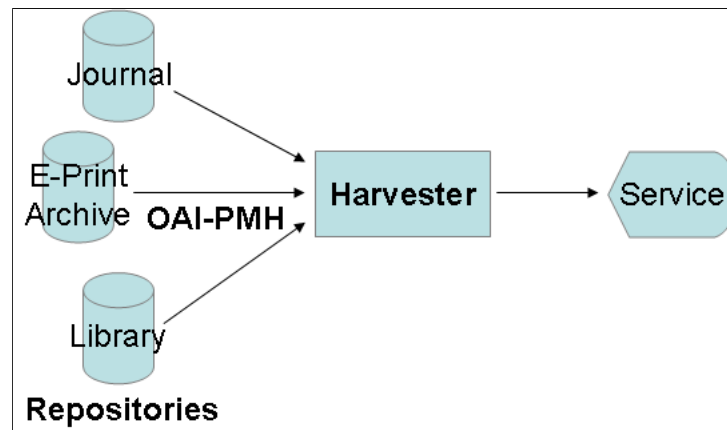
**Figure 3.6:** Liu et al. (2002) proposed infrastructure for extending OAI-PMH

### 3.4 The Open Archives Initiative

“The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication . . .” (Lagoze and Van de Sompel, 2001). While the OAI has its origins in the open access community (Van de Sompel and Lagoze, 2000) the technical implementation isn’t predicated on open access – Van de Sompel and Lagoze (2002) states that “[in the last year] the OAI-PMH has emerged as a practical foundation for digital library interoperability.”

The *OAI Protocol for Metadata Harvesting* (OAI-PMH) allows distributed repositories of documents to be harvested to form a single, aggregated collection (Figure 3.7). The purpose of the OAI-PMH is to allow repositories to expose, as easily as possible, their collections to service providers. By minimising the barrier to interoperability OAI-PMH aims to achieve widespread adoption, hence establish an environment where services can more easily access material to build collections from.

Van de Sompel and Lagoze (2002) point to OAI-PMH’s general acceptance as due to being “intentionally low-barrier, exploiting widely deployed web technologies such as HTTP and XML. It builds on many years of metadata practice, leveraging the development of a lingua franca metadata vocabulary



**Figure 3.7:** High-level OAI-PMH data flow-chart.

in the Dublin Core Metadata Initiative. It accommodates a number of community and domain-specific extensions such as the co-existence of multiple domain-specific metadata vocabularies, collection descriptions, and resource organization schemes.”

A repository contains records that describe items in its collection. An item is anything that can be described by metadata *e.g.* a published paper described by bibliographic data (title, author etc). An item has a unique identifier and is described by one or more metadata formats. Each metadata format is identified by a repository-unique metadata prefix. A metadata record contains the metadata about the item in a given format, that is a standardised way of marking-up that data in XML. (Identifier and metadata prefix uniqueness is only required within the scope of the repository; therefore different repositories may use the same identifier to describe entirely different items.)

The identifiers used by the repository allow harvesters to later request updates for a specific item only. The only other requirement for identifiers is that they conform to the URI format (Uniform Resource Identifiers), which allows the item identifier to be encoded as a text string (NB web addresses are URIs hence can be used as OAI identifiers). To date most repositories have re-used existing identifiers for their OAI items *e.g.* arXiv pre-pended `oai:arXiv.org:` to the existing repository-specific identifiers for their records to make them into valid OAI identifiers, hence usable in their

OAI-PMH interface. However, re-using existing identifier schemes reduces the repository's flexibility to control how services display their records *e.g.* by using a separate OAI set of identifiers repositories can combine multiple internal records together or instruct services to delete OAI records without invalidating a live internal identifier.

The ability to tie multiple metadata records to a single item allows OAI to transport parallel metadata formats, allowing for communities to have specialised metadata but still allowing generic lowest-common-denominator services to aggregate heterogeneous collections. Because the OAI-PMH is XML-based any metadata shared using OAI needs to be encoded in XML. All OAI-compliant repositories are required to at least support the Dublin Core metadata format. To expose other metadata formats a repository defines a repository-unique prefix, which is then used by a harvester by specifying that prefix when harvesting records. A repository need not expose every item in every supported metadata format, allowing a repository to contain different types of items (books, paintings etc.) with metadata appropriate to each type.

### 3.4.1 Dublin Core

Dublin Core consists of 15 terms (metadata fields), each of which may have zero or more text values. While some Dublin Core terms can be used to describe almost anything (identifier, title, description in particular) it best describes bibliographic objects – things written by someone, distributed as an electronic document in a particular format, and possible derived or linked to other documents.

The 15 terms ([Table 3.1](#)) can be refined using qualifiers to more specific meanings *e.g.* 'abstract' (as-in a research paper abstract) is a refinement of 'description'. However, the OAI-PMH Dublin Core metadata format uses only the 15 unqualified terms.



**Table 3.1:** Dublin Core Metadata Element Set

| Term        | Description  |
|-------------|--|
| contributor | An entity responsible for making contributions to the content of the resource. |
| coverage    | The extent or scope of the content of the resource.                            |
| creator     | An entity primarily responsible for making the content of the resource.        |
| date        | A date associated with an event in the life cycle of the resource.             |
| description | An account of the content of the resource.                                     |
| format      | The physical or digital manifestation of the resource.                         |
| identifier  | An unambiguous reference to the resource within a given context.               |
| language    | A language of the intellectual content of the resource.                        |
| publisher   | An entity responsible for making the resource available.                       |
| relation    | A reference to a related resource.   |
| rights      | Information about rights held in and over the resource.                        |
| source      | A reference to a resource from which the present resource is derived.          |
| subject     | The topic of the content of the resource.                                      |
| title       | A name given to the resource.  |
| type        | The nature or genre of the content of the resource.                            |

### 3.4.2 Metadata Semantic Problems

The successful use of metadata harvested from repositories using the OAI-PMH relies upon a shared understanding between repository and service provider on what the metadata means. From a service provider's perspective this either requires repositories to have a common interpretation of metadata formats or to have systems in place that can normalise the records harvested from multiple repositories into a consistent aggregated collection.

Regardless of how detailed a format may be invariably ambiguity arises when a standard is exposed to real-world applications. In building services based on OAI Dublin Core (as the most widely implemented format) the inconsistencies between repository implementations have surfaced and, when it comes to building more advanced OAI-based services, additional work has been required to normalise.

The Dublin Core 'date' field has been used to contain the date of publication<sup>8</sup>, the date the record was created in the repository<sup>9</sup> or both in one case<sup>10</sup>. Inconsistent use of dates makes it difficult to provide an aggregated list of records ordered by their creation date (*vs.* when they were harvested by the service).

To perform autonomous citation linking requires access to the full-text as well as the metadata. The Dublin Core 'identifier' term should be used to give the URL of the digital object (or perhaps objects if there is more than one format). However, most OAI repositories have opted to put only the 'jump-off' page URL into the 'identifier' field. While a user may be able to easily navigate from the jump-off page to the full-text, it has proved troublesome for me to run a service that has to either scrape the web page to identify the location of the full-text files or to create rules to translate from the record's identifier to the full-text location.

---

<sup>8</sup>Default in GNU Eprints *e.g.* <http://eprints.aktors.org/>

<sup>9</sup>*e.g.* arXiv.org <http://arXiv.org/>

<sup>10</sup>BORA-UiB: Bergen Open Research Archive <https://bora.uib.no/>

## 3.5 OpenURL

To analyse citation data requires having a citation database. Building a citation index requires a collection of research papers, the ability to extract their references and citations and linking those references, citations and papers together. The ISI Web of Science is unique in that its central purpose is a citation index. However, citation linking, providing hyperlinks for citations to allow easy navigation, is widely implemented by digital library systems. In this section I outline two complementary systems – OpenURL and DOI (run by CrossRef) – that support citation linking for existing collections of research papers. These differ from the bespoke citation indexes outlined so far, in that they aim to provide citation linking across providers, and across heterogeneous collections.

### 3.5.1 Persistent Linking using OpenURL

While the capability of linking is well understood, and widely implemented, being able to reliably and persistently provide links to the location of an item has proved difficult. Links to items by location (*e.g.* a web page URL) fail when the location changes – resulting in ‘broken links’ ([Markwell and Brooks, 2002](#)) that waste users’ time and degrade the quality of the service.

Maintaining accurate location-based links by hand is an impossible task due to the rate of decay of web objects: [Koehler \(2004\)](#) gives a range of half-lives for web-based material ranging from 2 years for ‘random web pages’ to 4 years for ‘computer science citations’ (approximately one link failing every two weeks). In contrast Koehler calculated a half-life of 24.5 years for objects in public digital library web sites (based on [Nelson and Allen, 2002](#)), even though many of those objects changed location (*e.g.* ResearchIndex changed its internal linking structure breaking external links). The OpenURL framework provides a more reliable means to link between objects by embedding descriptive metadata into objects, making linking to an object independent of the current location of that object.

### 3.5.2 Contextual Linking using OpenURL and SFX

The motivation for creating OpenURL was not, however, persistent linking. Van de Sompel and Beit-Arie (2001) pointed out that “established linking frameworks provide service-links that fail to take into account the context of the user who follows a link” and that such frameworks are “narrowly focused, both regarding the types of extended services that are being provided as well as regarding the action radius of those links”. In other words existing links between information providers were not sensitive to the context of the user (*i.e.* what subscriptions the user may have) and what a link delivers is dependent upon the targeted provider, not what the user wants (*e.g.* to retrieve the full-text).

**Table 3.2:** Example KEV-encoded OpenURL

|                            |  |
|----------------------------|--|
| Resolver URL               | <code>http://resolver.my.org/openurl?</code>   |
| Bibliographic search query | <code>ctx_ver=Z39.88-2004&amp;<br/>rft_id=info:doi/10.1126/science.275.5304.1320&amp;<br/>rft_val_fmt=info:ofi/fmt:kev:mtx:journal&amp;<br/>rft.genre=article&amp;<br/>rft.jtitle=Science&amp;<br/>rft.aulast=Bergelson&amp;<br/>rft.date=1997&amp; rft.volume=275...</code> |

An OpenURL is an encoding of bibliographic metadata (*e.g.* see table 3.2) that can be passed to an *OpenURL Resolver* which uses that metadata to locate the cited object. The metadata is either in Key-Encoded Values (KEVs) or in XML. The metadata for a journal paper might include the authors’ names, journal title, volume and pagination, and persistent identifier (*e.g.* DOI). While OpenURL is an open standard for the transport of bibliographic metadata (ANSI, 2004), the framework for linking (how an OpenURL resolver resolves OpenURL links) is implementation dependent.

Van de Sompel and Beit-Arie (2001) developed an ‘SFX server’ (an OpenURL resolver), that was later turned into a commercial product by Ex Libris<sup>11</sup> Walker (2003). SFX is a high-level infrastructure for implementing context-sensitive, dynamic linking to intellectual works (*e.g.* a research

<sup>11</sup>Ex Libris <http://www.exlibrisgroup.com/sfx.htm>

paper). SFX is based on a two-layer model of a ‘metadata layer’ and a ‘linking layer’. The user traverses these layers by following OpenURLs from the metadata layer to the linking layer, and then following an absolute URL generated from the OpenURL at the linking layer to the metadata layer.

How the user is directed in the OpenURL framework is dependent on the user’s *context*. An OpenURL ContextObject contains six entities: the *requester* is the identity of the user, the *referringEntity* is the item the user is linked from, the *referrer* is the service that generated the link, the *resolver* translates the *referent* (the item linked to) into a location accessible by the user and lastly the *serviceType* represents the type of request the user is making (*e.g.* for the full-text).

For example if a user in a college is following an OpenURL link to a journal paper the resolver service could use that user’s context (that they are a member of that institution) to automatically direct them to a subscribed copy or, if the college doesn’t have a subscription, to provide a link to the library’s copy-request service. An OpenURL resolver service could allow that user to customise their ‘serviceType’ settings to *e.g.* always show the abstract for a full-text, rather than linking directly to a PDF copy.

In order for an OpenURL resolver to resolve an OpenURL it needs to have a database of metadata records to match the OpenURL against. How the resolver builds this database is not defined within the OpenURL standard. One possible solution is to use the OAI-PMH (see 3.4, page 39) to transfer bibliographic records from OpenURL targets (the services that are linked to).

OpenURL is useful for some aspects of citation linking but not others. OpenURL provides a framework to link citations within a hyperlinked environment (*i.e.* the web). OpenURL is also a useful standard for encoding and transferring bibliographic metadata either encoded as URLs (Key-Encoded Values – KEVs) or in XML. The drawback of OpenURL is it unclear whether an OpenURL is resolvable – it is dependent on the resolver service. This is particularly an issue for autonomous systems – such as Citebase Search – that might only be able to generate partial references

(*e.g.* missing the journal title).

### 3.5.3 Digital Object Identifier

The purpose of reference linking is to resolve human-structured references to either globally or, within a closed system, unique identifiers. [Atkins et al. \(2000\)](#) describes the Digital Object Identifier, or DOI, developed by publishers to globally identify published works. DOI is a hierarchical identifier that allows publishers to purchase a unique, and persistent, identifier space with which to identify the literature they control. The DOI linking infrastructure is based on querying a database of references, along with their matching DOIs, to resolve a reference to its unique DOI. The DOI can then be used to build a web link to a publisher's web site, where the referenced paper can be retrieved.

The DOI system allows a publisher to maintain globally unique identifiers for electronic material they produce. The theory is that DOI's will be more stable than URLs, for *e.g.* if a publisher sells their material to another publisher the DOI system can re-map the DOI to the URL's of the new publisher.

DOI relies on a central, managed index of DOI objects to resolve DOI's to the identified object.

CrossRef<sup>12</sup> is the organisation that implements the DOI system. In essence CrossRef is a registry of electronic resources, with each resource containing a bibliographic description (metadata) and a DOI to uniquely identify it within the DOI system.

The relevance of CrossRef to citation linking is that it provides an authority that can be queried when searching for a cited paper, and by using the returned DOI a service can provide a link to the full-text of the cited paper. Therefore, when a electronic publisher adds an paper to their site as well as registering the paper itself, the publisher can query CrossRef for every

---

<sup>12</sup>CrossRef <http://www.crossref.org/>

reference contained in the paper, adding DOI links for any found in the CrossRef database. As most references contain only limited bibliographic data (journal, volume, author, etc.) the CrossRef system allows queries using only a partial record to discover and return the DOI along with the canonical bibliographic record.

Since mid 2004 CrossRef has extended citation linking to include forward linking (*i.e.* citation analysis). The forward linking service allows a participating service to query CrossRef for all citations to a given DOI. Having got a list of citing papers the service can present that list to the user. As forward linking only allows per-paper querying, it can't easily be used to provide other citation analyses *e.g.* co-citation and citation-coupling.

CrossRef is an independent organisation whose policies are determined by a steering group made up of its members. To access CrossRef's data, or to be able to submit new bibliographic records, requires paying a membership fee. The fees in turn cover the cost of development, supporting the DOI service, etc. While CrossRef (and the DOI system) is undoubtedly successful in the commercial publishing industry, the cost of membership is prohibitive to using their data in an open access citation database. There are also considerable costs associated – on the part of publishers – in formatting reference data for submission to CrossRef (this is the main cost associated with building an open access citation index).

## 3.6 Conclusion

Both commercial and free tools have been developed that provide users with citation links. These tools use heuristics (rule-based parsing) to extract references from full-texts, parse them into structured bibliographic data and link them to build a citation database. End-user services provide a number of mechanisms for users to navigate around the literature using citation links, including (*e.g.* ScienceDirect) linking to a number of different services depending on where the cited paper is located.

The OpCit project has demonstrated building a service that harvests full-texts from open access repositories, citation links them and re-exposes that citation data in a form usable by others. While interoperability between OAI-based services is still at the early stages – being based primarily on Dublin Core and on a two-party, repository-service based infrastructure – there is a vision for a more federated approach to constructing services.

In this thesis I use citation data generated from open access papers using techniques developed in the OpCit project. I have also used data from the ISI Web of Science to compare open access papers against non-open access papers. This data enables the *bibliometric* analysis of open access material.



# Chapter 4

## Bibliometrics

Bibliometrics is the quantitative analysis of publication patterns within a given field or body of literature. Two commonly cited laws of bibliometrics are Lotka and Bradford. Lotka's and Bradford's laws describe respectively the frequency distribution of papers by author and the frequency distribution of papers in journals. Bradford's law is useful for managing journal purchasing decisions, as it gives an idea of how many journals are required to cover a given percentage of the total literature for a subject. Zipf's law can be used to describe the distribution of research papers by citation impact, even though it originated from the analysis of the word-frequency distribution of words in English.

The Zipf, Lotka and Bradford distributions are highly skewed – which, because most statistical methods are based on a gaussian distribution, makes analysing research literature more troublesome. Most analyses in this thesis look at the citation and usage (web download) distribution per paper, which conform to Zipf.

The second half of this chapter describes the *Journal Impact Factor* (JIF). The JIF was devised by Eugene Garfield as a quantitative measure of the importance of a journal, normalised by the journal's size (as we know from Bradford's law journals vary greatly in the number of contained papers). The JIF has gained in prominence both because of the pressure of library

cancellations, where libraries may use the JIF to determine less ‘important’ journals to cancel first, but also because it has been used in evaluating authors: the higher the JIF of the journal an author has published in, the higher ‘quality’ that author is.

## 4.1 Bibliometric Techniques and Laws

### 4.1.1 Zipf’s Law

[Zipf \(1949\)](#) demonstrated his law using English text. When the frequency of each word is counted, and the result rank-ordered by the frequency, the rank multiplied by the frequency is roughly constant. Zipf’s law is defined as

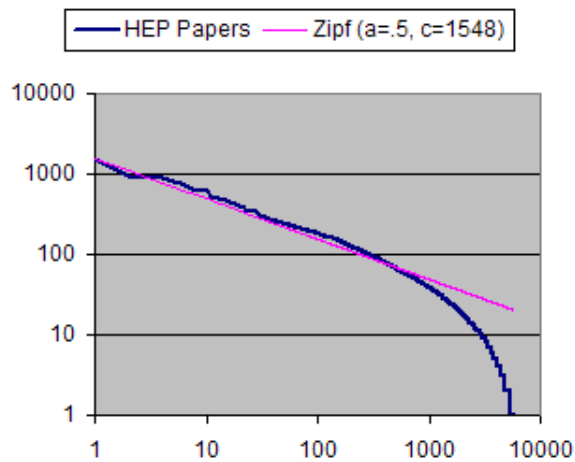
$$P_n \sim 1/n^a \quad (4.1)$$

where  $P_n$  is the frequency of occurrence of the  $n^{th}$  ranked item and  $a$  is close to 1. When plotted on double-logarithmic axis (where  $x$  is the logarithm of the rank and  $y$  the logarithm of the frequency) a Zipfian distribution is a straight line.

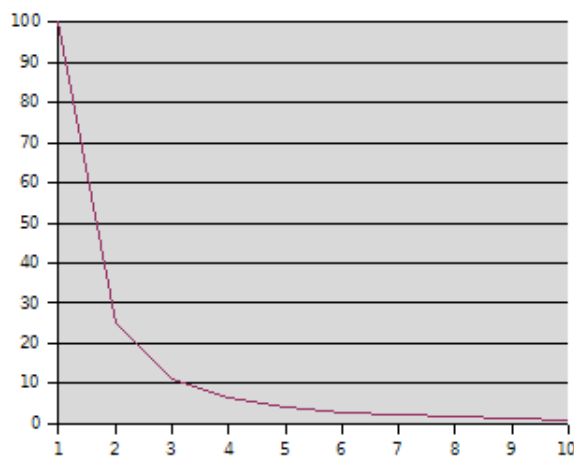
The distribution of papers by citation impact adheres to the Zipf distribution ([Redner, 1998](#)). [Figure 4.1](#) shows a sample set of papers from Citebase Search (High Energy Physics papers deposited in 1996) plotted on double-logarithmic axis, along with an estimated line of best-fit for Zipf ( $a = .5$ ). Many other naturally occurring distributions have been found to obey this law *e.g.* the sizes of cities in a country.

### 4.1.2 Lotka’s Law

“Lotka’s Law describes the frequency of publication by authors in a given field. It states that “... the number (of authors) making  $n$  contributions is about  $1/n^2$  of those making one; and



**Figure 4.1:** Papers rank-ordered by citation impact and a Zipfian distribution (double-logarithmic axis)



**Figure 4.2:** Lotka's Law distribution (1 contribution = 100)

the proportion of all contributors, that make a single contribution, is about 60 percent” (Lotka 1926, cited in [Potter 1988](#)). This means that out of all the authors in a given field, 60 percent will have just one publication, and 15 percent will have two publications ( $1/2^2$  times .60). 7 percent of authors will have three publications ( $1/3^2$  times .60), and so on.”

*‘Bibliometrics’ at the University of Texas*<sup>1</sup>

Lotka is the same distribution as Zipf but with an exponent  $a = 2$ . A more

<sup>1</sup>Description of various bibliometrics laws <http://www.gslis.utexas.edu/~palmquis/courses/biblio.html>

recent study by [López-Muñoz et al. \(2003\)](#) found 70% of authors with a single paper, 14% with two, 6% with three *etc.*, suggesting Lotka still holds true today.

### 4.1.3 Bradford's Law

Bradford's law describes the distribution of papers between journals. If all journals in a field are rank-ordered by the number of contained papers, and then divided into three groups containing equal numbers of papers, the first group of will contain  $n$  journals, the second group will contain  $n^2$  journals and the third group  $n^3$  journals (described as being  $1 : n : n^2$ ). This is useful to keep in mind when comparing bibliographic services that aggregate journals – assuming a service indexes the largest journals first, as the number of journals is increased so an ever decreasing increase in the proportion of papers is achieved.

### 4.1.4 Bibliographic Coupling and Co-Citation

The descriptions used here are a summary of ([Kampa, 2002](#), Chapter 5.5).

Bibliographic coupling, as first described by [Kessler \(1963\)](#), relates papers by their reference lists. Authors cite works that support their argument, or are the background to their work. Hence, two papers that cite the same works are likely to at least share a common background and will be within a related subject. The greater the number of shared references between papers, the more likely the papers are on the same topic.

Co-citation is a similar metric to bibliographic coupling but relates papers not by their references but by citations. Two papers that are cited in the same reference list are likely to be related. The more often those two papers are cited together the stronger the relationship between them.

Both coupling and co-citation metrics are used in Citebase Search (see [chapter 7](#)) and CiteSeer (see [subsection 3.3.2](#)) to provide citation navigation

to the user in addition to following linked references and citing papers. Bibliographic coupling is provided by the ISI Web of Science but called ‘Related Papers’ (see [subsection 3.2.1](#)).

## 4.2 Eugene Garfield and the Science Citation Index

Eugene Garfield (see [Garfield, 2002](#), home page) is considered the ‘grandfather’ of bibliometrics. Garfield set up the Science Citation Index (see [subsection 3.2.1](#), page 32) from which the Journal Impact Factor (JIF) is calculated – the *de facto* performance indicator for the publishing and research community. Journals compete for library subscriptions largely on their JIF ([Garfield, 1972](#)) and researchers may be evaluated by the JIF of the journals in which they have published (see next section for a more in depth look at the JIF).

[Garfield \(1955\)](#) put forward the idea of a citation index for the sciences as a way to improve the scholarly process (there had been such indexes for law reports for some time). A citation index is a list of papers along with the papers that cite them. Garfield also suggested the possibility of using citations as a measure of the impact of an article within its research field. Counting citations would provide a better indication of performance than the existing method of simply counting the number of publications an author had written.

[Garfield and Sher \(1963\)](#) presented results of research into the citation behaviour (“bibliometrics”) of research literature published in 1961. This found that when plotting citation frequency (the number of times something is cited, be that a paper, author or journal) that a small subset receive the majority of citations. For example, 60 of the 5000 journals studied accounted for 60% of all citations. This leaves the majority of papers receiving little or no citations after they are published.

Today the Science Citation Index covers around 8,700 journals ([ISI, 2004](#))

of an estimated 24,000 (Harnad et al., 2004). However, Garfield (1990) pointed out that “no matter how many journals are in the market, only a small fraction account for most of the articles that are published and cited in a given year.” Thomson ISI (2004) estimated “that a core of approximately 2,000 journals now accounts for about 85% of published articles and 95% of cited articles.”

### 4.2.1 The Impact Factor

The journal impact factor is the number of citations to a journal normalised by the number of papers in that journal (Equation 4.2 gives the mathematical definition). Garfield first put forward the idea of an impact factor in Garfield (1955) – “when one is trying to evaluate the significance of a particular work and its impact on the literature and thinking of the period ... such an ‘impact factor’ may be much more indicative than an absolute count of the number of a scientist’s publications.” While Garfield put forward the idea of an impact factor as a means of evaluating research it has been most widely used as a method of comparing the importance of journals.

Garfield wasn’t the first person to use citations as a quantitative measure of the importance of journals, Gross and Gross (1927) proposed counting citations as a means of collection management for journals (unlike the journal impact factor, Gross and Gross didn’t normalise for the number of papers in a journal). However, the increasing use of research evaluation has led to the use of the JIF to being the most widespread research metric used in evaluation.

$$I_j = \frac{C_{t-2}^t}{P_{t-2}^t} \quad (4.2)$$

where  $I_j$  is the impact factor of a journal  $j$ ,  $C$  is the total citations to papers in that journal over two preceding years and  $P$  is the total research papers published in that journal over two preceding years.

[Garfield \(2005\)](#) likens his creation to that of nuclear energy – “the impact factor is a mixed blessing. I expected it to be used constructively while recognizing that in the wrong hands it might be abused.” As Garfield acknowledges the JIF has been transposed from a measure of journal impact to a proxy of the impact of authors publishing in that journal. This has had real economic consequences for some researchers *e.g.* in Spain the use of the ISI JIF to award researchers’ bonuses is enshrined in law ([Jiménez-Contreras et al., 2002](#)).

The use of the journal impact factor to evaluate authors is, however, deeply flawed. [Seglen \(1994, 1997\)](#) argued against the use of the journal JIF as an evaluation tool for authors, as there is a huge range in the number of citations to papers within a journal (Seglen found for three biochemical journals that “15% of the [papers] account for 50% of the citations, and the most cited 50% of the [papers] account for 90% of the citations.”) In effect, an author of a low-impact paper gets the same rating as an author of a high-impact paper, as long as both are published in the same journal. While journals vary in the quality and type of papers they accept, within a journal individual papers will have a wide range of quality and hence impact.

The use of quantitative evaluation inevitably effects the subject of evaluation. If the evaluation system did not result in the subjects changing (hopefully improving) the evaluation isn’t having any effect. This evaluation pressure on journals has led to accusations of cheating the system to maximise a journal’s JIF. One mechanism a journal can use to increase their JIF is to encourage authors to cite papers previously published in that journal. [Fassoulaki et al. \(2000\)](#) found a very strong correlation of  $r = 0.899$  between the amount of journal ‘self-citation’ and its JIF. While there may be entirely reasonable motivations to do this *e.g.* to tie papers into the historical record of the journal, the net result is to increase that journal’s JIF.

## 4.3 Conclusion

Bibliometrics is the general umbrella field for the quantitative analysis of research publications. Reviewing the general laws of bibliometrics are useful as a background to the analysis of open access literature, because (to date) open access shares the same characteristics (the distribution of citations, authors and journals).

The Journal Impact Factor (JIF) was developed to compare journals' importance within a field (the more highly cited the journal, the more important it is). Because citation counts are recognised as a useful quantitative measure of importance, using citation data to evaluate the effect of open access is an obvious step (by comparing the citation impact of open access *vs.* non-open access papers). But the limitations of the JIF must also be kept in mind when drawing conclusions from any citation counting based comparison: that citations are highly skewed, 'self-citation' can distort results and that citations are essentially a measure of popularity and not necessarily of quality.

In order to perform bibliometrics at all requires access to bibliometric data. The Open Archives Initiative ([section 3.4](#)) provides me with metadata (authors, title etc.) for open access papers but in order to get citation data for these papers requires building a citation database, which is discussed in the next chapter and in Citebase Search ([chapter 7](#)).



## Chapter 5

# An Analysis of OAI Repositories and Harvesting Support

### 5.1 Introduction

Distributed systems are advantageous because they share the costs across many providers, improve reliability through removing single points of failure and, for many applications, improve performance by distributing load across multiple providers. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a distributed system – many data providers manage the acquisition and cataloguing of resources, providing a common interface for many services to harvest and aggregate those resources. But some problems in harvesting data providers have been encountered by OAI service providers.

Given the global nature of OAI-PMH – and its very low implementation cost – many OAI data providers are on low-bandwidth networks or exhibit minor errors in their ‘home-brew’ implementation. Some very large collections of material exist that present a very large amount of data to re-harvest should an error occur during harvesting.

Two problems encountered while harvesting from OAI data providers are errors in text data (character encoding issues), which causes the XML-based responses to not be parseable by XML parsing tools, and not correctly implementing the required parts of the OAI-PMH protocol (*e.g.* only accepting the optional seconds-based timestamps, when OAI-PMH requires at least support for day-based resolution).

To help with these challenges I wrote a tool that harvests records from data providers, handles OAI-implementation errors and stores the record metadata in a database for use by other services. This tool is called ‘Celestial’.

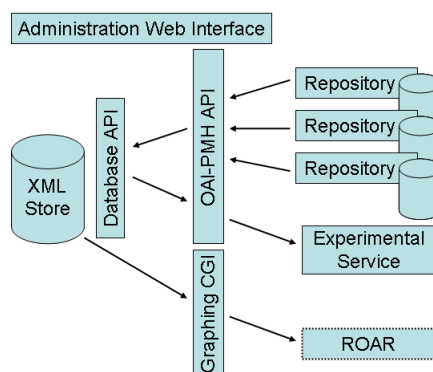
Celestial does not help to resolve the semantic ambiguity associated with Dublin Core metadata (see [subsection 3.4.2](#)) – or other problems with metadata interpretation. The purpose of Celestial is to at least provide a consistent (correct) mechanism to obtain metadata, by abstracting over the wide range of OAI-PMH implementations.

## 5.2 Celestial Architecture

Celestial consists of a MySQL database, a harvesting tool, an OAI-PMH web interface and an administrative interface. The data flow around Celestial is shown in [Figure 5.1](#). OAI records are harvested from OAI repositories using the OAI-PMH API (that handles errors – see later in this chapter) and stored in the database as raw XML. Celestial does not aggregate records from multiple repositories together, instead each harvested repository is stored and exposed from separate OAI baseURLs. Each baseURL consists of Celestial’s OAI interface (<http://celestial.eprints.org/cgi-bin/oaia2>) followed by the name of the repository ([arXiv.org](#)), *e.g.* to get the `Identify` response for the arXiv mirror a service requests

<http://celestial.eprints.org/cgi-bin/oaia2/arXiv.org?verb=Identify>.

The motivation for mirroring content in Celestial (*vs.* acting as a proxy) is twofold: firstly it reduces repeated accesses to the source repository because the



**Figure 5.1:** Celestial's Architecture

mirrored copy is autonomous and secondly Celestial is designed to handle very slow remote sites that might otherwise block a service's harvesting. Having a copy of the data also allows analysis of the OAI records (*e.g.* to supply records graphs to ROAR – the 'Graphing CGI' in [Figure 5.1](#)).

Most of Celestial's functionality is inside the OAI-PMH API (a separate module, usable by other OAI-based services). The core of Celestial consists of an XML store that adds a datestamp and unique (internal) identifier to each harvested record. When a service harvests from Celestial it typically asks for any records changed since its last visit (as it would from a normal data provider). Querying the database for all of the records with a datestamp more recent than the date given by the service returns all of the new records. Celestial uses the OAI-PMH partial-listing feature to break up the list of matching records into 100-record long 'chunks' (so if a data provider doesn't use flow-control Celestial makes response sizes more manageable). To get the next chunk a harvester uses a 'resumption token'.

The resumption token returned by Celestial consists of the datestamp and the internal identifier of the last record in the chunk. When a service requests the next chunk Celestial performs the same date query but only for those records with an identifier number higher than the previous record. MySQL can very quickly return records starting with a given datestamp and identifier (by using an index over both).

590 repositories are registered in Celestial<sup>1</sup>, of which 562 have had some records harvested (NB 2 of those 562 are flagged as "permanently locked" – they have

<sup>1</sup>As of 2006-04-01, see <http://celestial.eprints.org/cgi-bin/status>

**Table 5.1:** Top ten most widely implemented metadata formats in Celestial-registered repositories

| Prefix      | Namespace   | Count |
|-------------|---|-------|
| oai_dc      | <a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>                                       | 489   |
| oai_rfc1807 | <a href="http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt">http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt</a> | 61    |
| oai_marc    | <a href="http://www.openarchives.org/OAI/1.1/oai_marc">http://www.openarchives.org/OAI/1.1/oai_marc</a>                                     | 61    |
| oai_etdms   | <a href="http://www.ndltd.org/standards/metadata/etdms/1.0/">http://www.ndltd.org/standards/metadata/etdms/1.0/</a>                         | 59    |
| marc21      | <a href="http://www.loc.gov/MARC21/slim">http://www.loc.gov/MARC21/slim</a>   | 48    |
| oai_dc      | <a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>   | 37    |
| oai_dc      | <a href="http://purl.org/dc/elements/2.0/">http://purl.org/dc/elements/2.0/</a>   | 27    |
| mods        | <a href="http://www.loc.gov/mods/v3">http://www.loc.gov/mods/v3</a>   | 17    |
| marc21      | <a href="http://www.loc.gov/MARC21/slim/">http://www.loc.gov/MARC21/slim/</a>   | 16    |
| epicur      | <a href="http://www.persistent-identifier.de/xepicur/version1.0/">http://www.persistent-identifier.de/xepicur/version1.0/</a>               | 13    |

failed for long enough to no longer be harvested). Celestial attempts to harvest every metadata format supported by each repository, which has resulted in 953 sets of metadata in approximately 60 distinct metadata formats. Table 5.1 gives a breakdown of the top ten most widely implemented metadata formats for Celestial-registered repositories. Several formats are associated with different XML namespaces (the globally unique way to refer to an XML document format), even though they are same underlying format (*e.g.* Dublin Core from OAI and Dublin Core from the Dublin Core Metadata Initiative<sup>2</sup>).

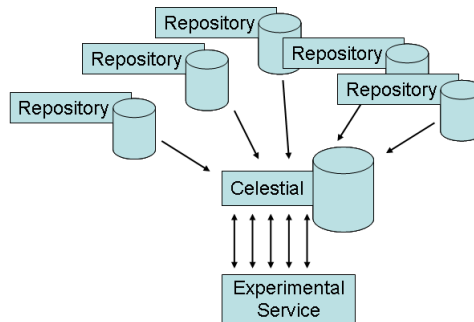
### 5.3 Reducing Repeated Requests

When building an experimental service using OAI it is sometimes necessary to wipe the database and re-harvest the raw metadata (*e.g.* where a process is modified that may effect the entire collection). For arXiv this represents some 400,000 Dublin Core records (about 1GB of XML). arXiv enforces a 60-second wait per 100 records, therefore to harvest 400,000 records takes about three days (assuming no other problems). Celestial operates as a local copy of arXiv's

<sup>2</sup>Dublin Core Metadata Initiative (DCMI) <http://dublincore.org/>

metadata that avoids asking arXiv for the same metadata more than once. Because Celestial stores the raw XML data it is possible to always retrieve the original metadata, whereas any more complex storage might introduce a bug, necessitating re-harvesting the entire collection. This is particularly an issue for Dublin Core where some fields may have an implicit order (*e.g.* lists of author names or the association between a format and a URL) that might be lost if the Dublin Core was parsed and stored as fields rather than XML.

Celestial is intended to cope with unreliable source repositories. Celestial requests updated records every day and only gives up if no successful harvest has been achieved in two weeks of trying. Each HTTP request to the server is retried up to three times in the event of an error.



**Figure 5.2:** Celestial supports experimental services by avoiding the need to repeatedly harvest source repositories.

## 5.4 Abstracting Multiple OAI Protocol Versions

Celestial supports harvesting from OAI-PMH version 1.0, 1.1, 2.0 and static OAI interfaces. All versions of the OAI protocol share the same six verbs and data model, which has allowed me to create an API that provides a common interface to Celestial, regardless of the version of the repository.

[Open Archives Initiative \(2002\)](#) provide a complete breakdown of the changes between version 1.1 and version 2.0. The changes that the API abstracts are: errors in version 1.x (1.0 and 1.1) are returned as HTTP errors and in 2.0 as XML documents, the `ListIdentifiers` command was changed to take a `metadataPrefix` argument in 2.0 (so it matches the syntax of

`ListMetadataFormats`) and the OAI-PMH document structure was changed between 1.x and 2.0: the root element's name was changed from the requested verb to 'OAI-PMH', the set membership of a record was added to record headers and `ListIdentifiers` returns the record headers (not just the record identifier).

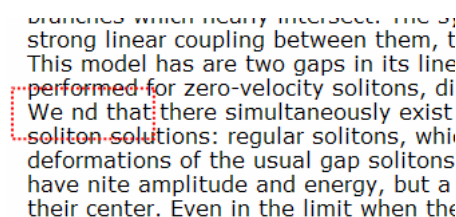
OAI static repositories use a single XML document to store the entire collection. As a static repository is a single file – and not a CGI interface – an initial request is always made by the API to the base URL to determine whether the repository is a static repository or a CGI interface (a CGI interface will respond with an error if no verb is given). If a static repository is encountered the XML document is cached and subsequent requests are wrapped around that cached document.

The API identifies the OAI-PMH version by checking the XML schema given in the root XML element (XML schemas describe the structure of an XML document). A header module checks whether the root element's name is the OAI verb (in version 1.x) or `OAI-PMH` (version 2). OAI errors are either returned as an HTTP error code (version 1) or an XML structure (version 2). The API wraps version 1 HTTP errors into version 2 XML structures. Any other errors that can occur, *e.g.* XML parsing problems, are also expressed as version 2 errors. The API converts version 1.x responses to 2.0 by *e.g.* wrapping version 1.x `ListIdentifiers` record identifiers into a full header object.

The OAI static repository proposal envisioned gateways “that makes Static Repositories harvestable through the OAI-PMH” ([Hochstenbach et al., 2003](#)), however the API used by Celestial fulfills this requirement as it can handle such repositories internally (so no gateway service is required).

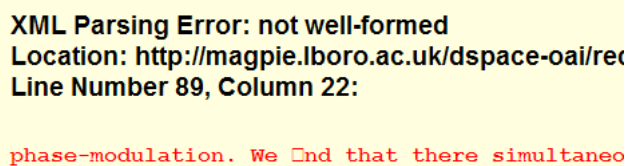
## 5.5 Correcting OAI Data Provider Errors

According to the [OAI-PMH \(2004\)](#) standard all responses to OAI-PMH requests must be well-formed XML instance documents encoded using the UTF-8 representation of Unicode. The nature of XML is that a single error invalidates the entire document. This can be a problem for OAI-compliant institutional repositories as they typically get their data from end-users, that can result in input that may not be obviously wrong in a web browser ([Figure 5.3](#)), but isn't compliant with XML ([Figure 5.4](#)).



branches which nearly intersect. The strong linear coupling between them, this model has are two gaps in its line performed for zero-velocity solitons, di We find that there simultaneously exist soliton solutions: regular solitons, which are deformations of the usual gap solitons, which have finite amplitude and energy, but a finite width at their center. Even in the limit when the

**Figure 5.3:** A bad character gets ignored by Internet Explorer when rendering a web page



**XML Parsing Error: not well-formed**  
**Location:** <http://magpie.lboro.ac.uk/dspace-oai/request>  
**Line Number 89, Column 22:**  
 phase-modulation. We find that there simultaneously exist soliton solutions: regular solitons, which are deformations of the usual gap solitons, which have finite amplitude and energy, but a finite width at their center. Even in the limit when the

**Figure 5.4:** In XML a bad character prevents the document from being parsed

Metadata values in an OAI response that have originated from sources that aren't in UTF-8 can cause problems if the necessary character encoding translation hasn't occurred. For example errors can occur when an author copies a section of text from a Microsoft Word Document into a web page form that contains quotes: unless the web browser correctly identifies the pasted character encoding the quotes can get translated by the user's web browser to the `latin-1` code point, whereas the repository thinks it is getting UTF-8 (see [Table 5.2](#) for comparison of UTF-8 and `latin-1`). When the repository exports that data it ends up creating badly encoded XML documents due to embedding the badly encoded character data (*e.g.* DSpace at Loughborough<sup>3</sup> – see [Figure 5.4](#)).

In an ideal world OAI implementors would check for badly encoded data and remove or fix it before export. However there are many institutional repository software implementations (*e.g.* see [chapter 6](#)), not all of which check submitted data for conformity to UTF-8 and XML standards. To handle these badly behaved implementations Celestial includes support for dealing with incorrectly encoded XML documents.

`latin-1` and UTF-8 both use a sequence of single bytes to represent character data, but UTF-8 uses multiple bytes to represent characters higher than the 127th Unicode point (Unicode is a taxonomy of characters used in written languages).

<sup>3</sup>Citing an example is difficult given the ephemeral nature of bugs, however on 2006-04-04 this was still the case: [http://magpie.lboro.ac.uk/dspace-oai/request?verb=ListRecords&from=2006-01-13&metadataPrefix=oai\\_dc](http://magpie.lboro.ac.uk/dspace-oai/request?verb=ListRecords&from=2006-01-13&metadataPrefix=oai_dc)

**Table 5.2:** Character encoding in UTF-8 and latin-1

| Range       | Covers                               | UTF-8       | latin-1     |
|-------------|--------------------------------------|-------------|-------------|
| 0-32        | Control Characters                   | Single-Byte | Single-Byte |
| 33-127      | Numbers, Roman-Alphabet, Punctuation | Single-Byte | Single-Byte |
| 128-255     | European, Math, Line-Drawing         | Multi-Byte  | Single-Byte |
| 256-onwards | International, Symbols, Specials     | Multi-Byte  | Unsupported |

Because the characters of the roman-alphabet (*i.e.* western languages) are all below 127 the vast majority of western-language text has the same binary representation in `latin-1` or `UTF-8`. This means handling badly encoded characters resulting from a `latin-1` to `UTF-8` conversion is the minority case, hence removing those characters doesn't destroy too much of the existing text. In Celestial any bytes that are above 127 but are either not part of a correct `UTF-8` multi-byte encoding or aren't a valid Unicode character are replaced by '?'.

While `UTF-8` may include any character from 0-127 (with multi-byte sequences above that) XML also prohibits the use of most *control characters*. These control characters are in the range 0-32 and represent general cursor and serial communication commands (*e.g.* 'End of Transmission'). Celestial removes all control characters except for those allowed in XML: tab, new-line, carriage-return and space. (NB control characters in OAI are most likely the result of an incorrect encoding translation, rather than an intended use.)

Once Celestial has 'cleaned-up' the data from the OAI repository it can be passed to an XML parser. Given correct XML whether metadata can be successfully harvested depends on whether the repository has 1) provided the correct XML structure for OAI-PMH, and 2) has implemented the interface protocol correctly (*e.g.* to allow incremental harvesting).

The majority of repositories manage to implement the document structure and protocol correctly, partly because there are tools to test an OAI-PMH implementation *e.g.* the OAI-validator ([Warner, 2005](#)). A couple of repositories that haven't managed to implement the protocol correctly are: RePEc (see [subsection 3.3.1](#)) who – as of April 2006 – generated an error on resumption token by incorrectly requiring the *metadataPrefix* argument (the resumption token



argument should be exclusive) and BieSON<sup>4</sup> who would not accept a day-granularity datestamp (OAI-compliant repositories must support at least day-granularity, seconds granularity is optional). I modified Celestial to use the granularity given in the repository's *Identify* response that allowed BieSON to be harvested. To harvest RePEc would necessitate making an incorrect OAI request that, if RePEc fixed the bug, would cause more problems.

## 5.6 Analysing OAI Data Providers

The OAI-PMH is a standard for the incremental transfer of records from a repository (archive) to a service. Celestial is an OAI-PMH caching proxy; it harvests records from OAI-PMH compliant sites, stores them in a database as XML and then re-exposes those records through its own OAI-PMH interfaces (one per repository).

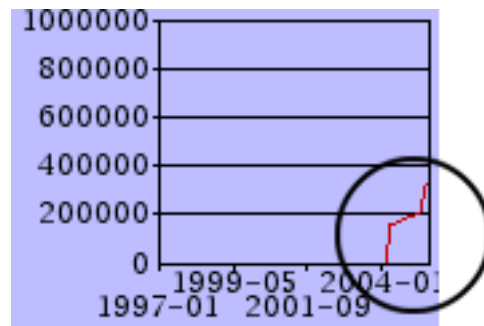
OAI-PMH records contain a unique identifier and a datestamp (the date the record was last modified or created, if the record has never been subsequently altered). The purpose of the datestamp is to allow incremental harvesting – an OAI-PMH service can request only those records that have been created or modified since the previous harvest.

The Registry of Open Access Repositories (ROAR – see next chapter) utilises Celestial's database of OAI records by extracting the OAI datestamps and presenting them as a cumulative graph of records over time. The intention is to track the growth in the number of records over time. The difficulty with tracking records over time is there is no consistent use of the 'date' Dublin Core field (see [subsection 3.4.2](#)), which means it can't be used in aggregations. The only other date available from an OAI record is the OAI datestamp (which is the date the OAI record was created or last updated). The drawback of using the OAI datestamp is that it is the date of the *record* and not necessarily the *resource*. This means it may not be usable for retrospective data (because OAI records can only have been created since the protocol was published), however the semantics of a datestamp mean that in incremental harvesting it serves the need to track

---

<sup>4</sup>BieSON - Bielefelder Server für Online-Publikationen (University of Bielefeld, GERMANY) <http://bieson.ub.uni-bielefeld.de/>

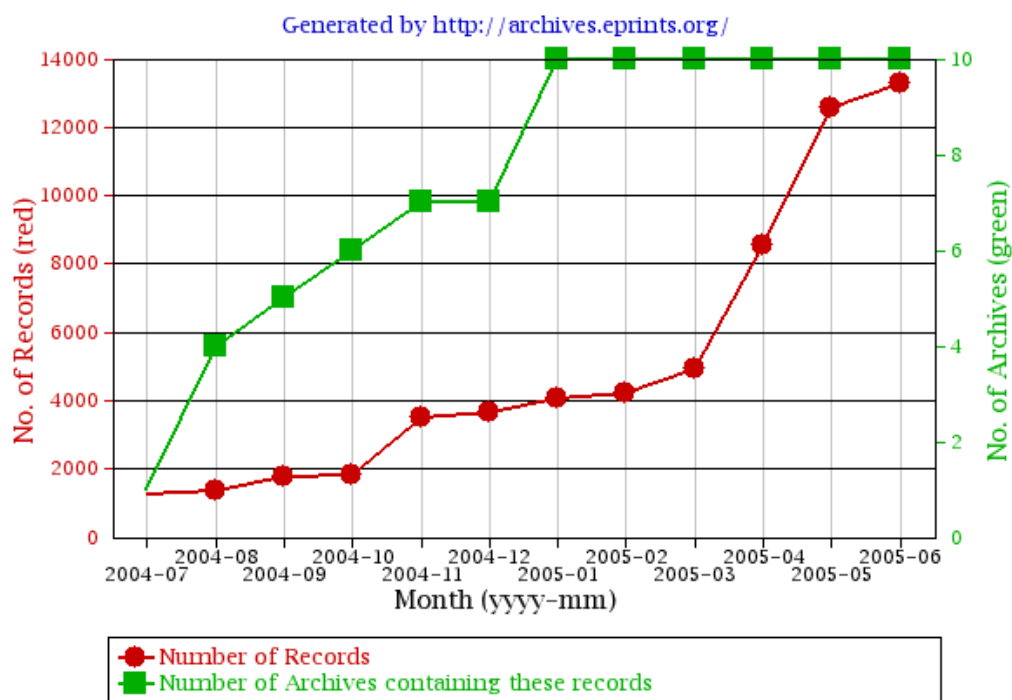
the growth of records (a record can't be created with an old datestamp nor should it have a datestamp in the future, both of which can apply to Dublin Core dates).



**Figure 5.5:** Thumbnail graph of arXiv's records generated by Celestial for ROAR, that shows two jumps in the number of records

**Figure 5.5** shows the records thumbnail graph for the arXiv (as linked to by ROAR), based on the datestamps from OAI records harvested from arXiv. As the datestamp is the OAI record datestamp, rather than when the digital item was created, many records may appear to be created simultaneously, whereas in reality a systematic modification was made to existing records (shown as the two jumps). For arXiv its OAI-PMH records were altered in late April 2004, hence causing every record's datestamp to change resulting in what at first appears to be 150,000 new records when those digital items have actually been steadily created since 1991.

The script used to generate the records thumbnail graphs accepts any number of OAI base URLs. This allows summary graphs to be created for any arbitrary collection of repositories *e.g.* repositories running a particular type of software, or from a given country. At the top of each record listing in ROAR is a 'Summary Graph' button that passes the URL from every listed repository to the script (repositories without a functioning OAI-PMH interface can't be counted and are ignored). The resulting graph contains two data series: the cumulative record count in red with circle points and the cumulative number of repositories in green with square points. The number of repositories is determined by taking the datestamp of the earliest record as the creation date for each repository (*e.g.* if an repository's earliest record datestamp is 2003-04-15 then that is when that repository will be added to the cumulative count). For any given month dividing the number of records by the number of repositories results in the mean records per repository for that month. **Figure 5.6** shows an example summary graph for repositories running the Bepress software.

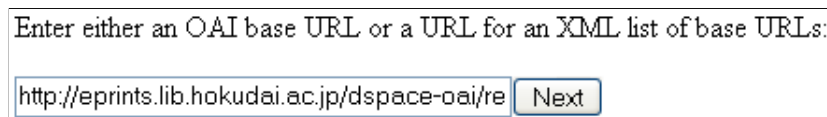
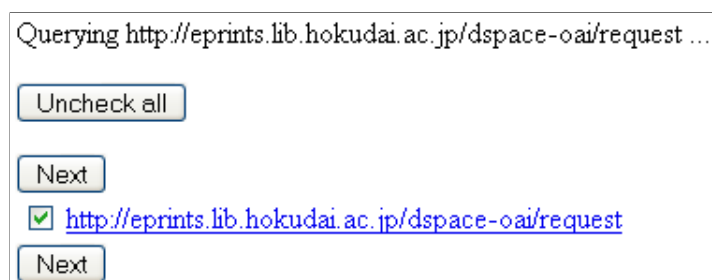


**Figure 5.6:** Celestial summary graph for Bepress-based repositories

When an entry in the ROAR has an OAI base URL specified, but has not yet been added to Celestial, the graphing script shows a ‘Not registered in Celestial’ notice. When the URL has been added to Celestial but not yet harvested the notice says ‘No successful harvest yet’. Once a successful harvest has been made the records graph is shown. If a harvest was made, but no record timestamps were found, a notice is given saying ‘No record dates found’. There are a multitude of reasons for a harvest to fail but the interface between the ROAR and Celestial is not advanced enough to provide more feedback to repository administrators (ROAR just links to Celestial). A repository administrator can use the Open Archives registry to test their interface to discover what the problem is (Celestial isn’t intended as a debugging tool) or check the Celestial status page for possible error messages.

## 5.7 Adding Repositories to Celestial

In order to produce records graphs in ROAR the repository must first be added to Celestial. Each repository is added using Celestial’s bespoke administration tool. The easiest way to add a new repository to Celestial is to use the import


**Figure 5.7:** Adding repositories to Celestial by URL

**Figure 5.8:** Selecting which URLs to add

feature. This takes the URL of an OAI baseURL or a web page listing of repository baseURLs (Figure 5.7), retrieves it, tests whether the URL itself is an OAI interface and parses the response for URLs. Any URLs found are presented as a list (Figure 5.8), that allows the administrator to select which URLs to test for being an OAI interface (this generic URL capture allows any web listing of repositories to be imported into Celestial). Clicking ‘Next’ attempts to retrieve the **Identify** response for each ticked URL and – if successful – the OAI interface is added to Celestial (Figure 5.9), to be harvested during the next scheduled harvest.

## 5.8 Conclusion

Celestial is a mirroring tool that normalises multiple OAI versions and handles several errors encountered in OAI implementations. Celestial provides analytical tools to ROAR (next chapter) and acts as a proxy for the experimental OAI

|                        |   |
|------------------------|---|
| <b>ProtocolVersion</b> | 2.0   |
| <b>RepositoryName</b>  | Hokkaido University collection of scholarly and academic papers |
| <b>adminEmail</b>      | repo@lib.hokudai.ac.jp  |

**Figure 5.9:** The URL is recognised as an OAI-PMH interface and ready for harvesting

service Citebase Search (see [chapter 7](#)).

As a publicly accessible service Celestial can be used by any service provider. In particular some users have used Celestial's copy of arXiv to avoid arXiv's throttling (which restricts requests to at most one per 60 seconds) *e.g.* [Ward \(2003\)](#).

## Chapter 6

# Quantifying Open Access in Institutional Repositories

In the previous chapter I described the Celestial tool which provides OAI record counts to the Registry of Open Access Repositories (ROAR) tool. ROAR is an index of research repositories across countries, disciplines and institutions (not just OAI-compliant repositories). Currently over six hundred repositories are registered, ranging in size from a few to millions of records. In this chapter I provide an overview of the current state of repositories of research-materials and attempt to determine how many records those repositories contain and how many of those records represent author self-archived, peer-reviewed research literature.

### 6.1 Introduction

The Registry of Open Access Repositories<sup>1</sup> (ROAR) was originally created to make it easier to monitor the uptake of the GNU EPrints software ([Gutteridge, 2002](#)). ROAR has since been extended to cover repositories running any software *e.g.* DSpace ([Smith, 2004b](#)), ETD-db ([Jones, 2004](#); [Virginia Tech, 2006](#)) or Fedora ([Payette and Lagoze, 1998](#); [Payette and Staples, 2002](#)). It also now has many entries that aren't 'institutional' or even 'repositories', but are still relevant to open access.

---

<sup>1</sup>Registry of Open Access Repositories <http://archives.eprints.org/>

As lead developers of GNU EPrints – software for creating OAI-compliant institutional repositories – the IAM Group at the University of Southampton needs to monitor the uptake of its software in order to support EPrints development, promotion and distribution. Assessing the success of open source software is difficult, as there are no sales to users. Instead, EPrints users are invited to register with the ROAR. Using the software-type field in ROAR we can easily count the number of EPrints-based repositories versus other softwares.

As well as facilitating open access through developing GNU EPrints the University of Southampton is promoting open access to the research literature. To encourage authors to self-archive ([Harnad, 1995, 2001b](#); [Pinfield, 2004](#)) we monitor the progress of research repositories using the ROAR: showing evidence for the uptake of open access is a strong motivator for other authors and institutions to provide open access to their own research (the ‘me-too’ argument). The type of repository software used is irrelevant to the goal of monitoring author self-archiving, however the lack of standard means to access the content of repositories presents problems.

Ideally we would like to know how many research papers are deposited in these repositories, compared to other digital materials (*e.g.* Powerpoint presentations). But this is currently not possible because repository implementations don’t provide a standard means to get only those OAI records that represent peer-reviewed, full-text research papers. An alternative approach to compare e-print repositories’ content against an authoritative list of peer-reviewed papers is beyond the scope of this project (*e.g.* by comparing a sample of records by hand against the ISI’s Web of Science). So while there is no doubt that there is an ever-increasing amount of open access material ([Morrison, 2006](#)), no study has yet provided a reliable figure for the proportion of institutional repository-based open access material that is peer-reviewed research papers.

### 6.1.1 Repositories *vs.* Archives

ar·chive

1. A place or collection containing records, documents, or other materials of historical interest. Often used in the plural: old land deeds in the municipal archives.

2. An archive for stored memories or information: the archive of the mind.

re·pos·i·to·ry

1. A place where things may be put for safekeeping.

*www.dictionary.com*

The terms repository and archive have a mixed use within the digital library and open access communities. In the library community an archive typically means providing long-term preservation of a collection (not necessarily to make the collection accessible to more users). In the open access community an e-print archive's purpose is to increase access – and hence use and impact – to research material by providing an author-supplied free, online version of the full-text. As [Stephen Pinfield and MacColl \(2002\)](#) points out the 'archive' in Open Archives Initiative (OAI) "refers primarily to the process of depositing of articles, rather than to their preservation."

The open access community originally focused on providing access to the peer-reviewed literature through subject-based repositories of e-prints (electronic pre- or post- prints of the published peer-reviewed papers). The weakness of a subject-based approach to e-print repositories is that there is no clear body that can take responsibility for the provision and ongoing support for an e-print repository service in every discipline. To address this issue the approach to the provision of open-access to e-prints has changed to an institutional focus. This institutional focus leverages a responsible body within the research institution – *e.g.* the library ([Crow, 2002](#)) – to establish and support a repository for the use of that institution's research staff: the Institutional Repository (IR). This can then be backed up by an institutional policy to encourage or mandate authors to deposit their work in the IR ([Harnad, 2006](#)).

Providing open access to e-prints is however not the only potential use or benefit of institutional repositories. As well as fulfilling the need of the institution to manage and maximise the impact of its research output, the IR can become a repository for teaching materials, databases and other forms of scholarly output – in essence the institution can capture the results of all scholarly activities ([Lynch, 2003](#)).



Part of the management of an IR is to provide a stable platform for researchers' work. A commitment to long-term preservation may or may not form part of an IR's policy *e.g.* digital preservation may not be needed at all if a version of the repository's content is archived by a responsible body through the legal deposit process for printed journals and books. However, there is widespread interest within the digital library community in taking advantage of IRs to provide long term access to and management of scholarly material produced by their faculty.

While in an ideal world – for open access proponents – the focus would continue to be on increasing access to that core peer-reviewed literature, with other research material and preservation being an added bonus, this mix of material in institutional repositories can lead to authors not seeing them as a valuable (additional) outlet for their research papers. So, while the ROAR and Celestial may count all objects equally my goal is to separate and count only those peer-reviewed research papers (to encourage repositories to focus on gathering high-impact research papers).

## 6.1.2 Existing and New Lists of Repositories

[Hitchcock \(2003\)](#) provides a 'metalist' of repository listings, covering a variety of open access sources. Many of these listings are generated from services *e.g.* OAIster<sup>2</sup>, rather than just registries like ROAR.

The OAI Registry of Data Providers<sup>3</sup> provides a list of sites with functioning OAI-PMH interfaces. Sites are added by submitting an OAI base URL (the location of the OAI interface); an automatic process then verifies the interface's compliance with the standard and, if it passes, the site can be added to the listing. The OAI Registry stores only the site's OAI base URL and a title. As few OAI repositories include collection-level descriptive metadata the potential for further analysis based on the OAI list is limited (*e.g.* sites can include the eprints schema in their **Identify** response, which provides a descriptive and rights date for the collection).

---

<sup>2</sup>OAIster collections listing

<http://oaister.umd.umich.edu/o/oaister/viewcolls.html>

<sup>3</sup>OAI Registered Data Providers <http://www.openrepository.org/Register/BrowseSites>

The OAI repository explorer tool<sup>4</sup> is intended as a tool for repository administrators to test and develop their OAI interfaces. It also has a registration facility that allows administrators – once their interface passes the registry’s automatic tests – to publish the URL of their OAI interface along with the title of their site. Like the OAI registry, the repository explorer tool provides no further information on the content of the registered repositories.

The UIUC registry<sup>5</sup> contains the largest number of entries of any repositories list (1084 at time of writing). The registry includes only OAI-compliant services but provides numerous analyses of the implementations and features of the OAI-PMH interfaces. For example, the user can locate repositories supporting particular metadata formats. Sites are added by sending an email to the administrator or are collected from other sources. Thomas G. Habing and Mischo (2004) describes a simple mechanism the UIUC registry uses for discovering OAI sites by using the Google query ‘allinurl:verb=Identify’. Although the UIUC registry provides many useful tools, it cannot indicate the number of (say) GNU EPrints sites, nor the number of records in each repository.

The OpenDOAR project<sup>6</sup> aims to “provide a comprehensive and authoritative list of academic open access research repositories for end-users who wish to find particular repositories or who wish to break down repositories by locale, content or other measures.” OpenDOAR is the sister project to the Directory of Open Access Journals<sup>7</sup>, which provides a listing of and some table of contents for open access journals. OpenDOAR currently contains 365 entries, however those entries are categorised by content type and subject (*i.e.* considerable effort has gone into checking and augmenting compared to other listings). Of those 365 entries, 228 were identified as containing ‘Articles’. OpenDOAR also provides listings of entries with ‘Conference papers’ (149) and ‘Pre-print journal articles’ (92), but it doesn’t provide a way to count the number of entries that contain any one of those types of content, nor what proportion of the content they represent.

---

<sup>4</sup>Repository Explorer by Hussein Suleman <http://purl.org/net/oai-explorer>

<sup>5</sup>UIUC Registry <http://gita.grainger.uiuc.edu/registry/>

<sup>6</sup>OpenDOAR <http://www.opendoar.org/>

<sup>7</sup>DOAJ <http://www.doaj.org/>

## 6.2 The Registry of Open Access Repositories

“We are promoting open access to the research literature pre- and post-peer-review through author self-archiving in institutional eprint repositories. Open access to research maximises research access and thereby also research impact, making research more productive and effective.

This registry has two functions: (1) to monitor overall growth in both the number and the contents of eprint repositories and (2) to maintain a list of GNU EPrints sites (the software Southampton University has designed to facilitate self-archiving).”

*The ROAR FAQ*

While there are now many lists of repositories available, no others yet provide the facilities we need and have implemented in ROAR: looking inside the repositories to determine at least the number of objects they contain. ROAR is an ongoing service, publicly accessible to anyone. Administrating ROAR requires identifying new repositories to add and encouraging administrators to register, removing ‘dead’ repositories and constructing tools to better analyse the data we collect.

### 6.2.1 Criteria for Inclusion in ROAR

As long as the number of repositories in ROAR is relatively small, there is no need to be too selective about what kinds of collections to include. The ROAR seeks new entries from two broad communities:

1. Services running established known IR software
2. Services with an OAI interface

These communities are separated into different ‘types’, based on the content included and who has contributed that content – the list of types was a subjective summary of the kinds of repositories encountered. The possible types are

‘Demonstration’, ‘e-Journals/Publications’, ‘e-Theses’, ‘Research Cross-Institution’, ‘Research Departmental/Institutional’, ‘Databases’ or ‘Other’. These types are not mutually exclusive – entries are categorised by where they best fit rather than by a strict criteria. For example a ‘Demonstration’ repository may be an institutional repository that has only a few records uploaded by one person – even though it may not be called a demonstration such a service is most likely an experiment.

A few entries come under the broad ‘Database’ type *e.g.* antbase.org (a taxonomic database of ant species). While such entries are added to the ROAR, database services are not actively invited to submit because databases are not central to the primary goal of promoting open access to research papers.

Many research departments have publication lists, from *e.g.* a single Word doc to searchable web databases, but if they do not have a clearly defined ‘service’ (either by having an OAI interface or being managed by IR software) it is difficult to usefully include these in the listing, because their content can’t be aggregated.

The ROAR includes services that are publication databases only (*i.e.* contain no full-texts or access-restricted ones). The service should be, or at least promote, open access, perhaps through the ‘keystroke strategy’ suggested by Carr and Harnad (2005): depositing all the full-texts, setting as many as possible to open access, and allowing end-users to generate emailed requests for secured eprints, to which the author can respond by emailing the eprint of the full text to the requester.

The purpose of the ROAR is to provide information beyond that covered by other lists. Autonomously harvesting entries leaves considerable work to be done by the ROAR moderators in filtering and augmenting what is typically just the OAI base URL located by the harvester. While the ROAR has the capability to retrieve URL’s from other services, it doesn’t regularly perform this process due to the time required to apply these criteria by hand. As the various publicly accessible registries mature hopefully standards for making the contained data more interoperable will be developed.

### 6.2.2 Removal from the Registry

The only grounds for removal is if a site has died, moved or duplicates an existing entry (as a result of moderator-error). I am unaware of any sites that have changed their policy on open access or type of content such that they should be removed from the listing: the criteria for inclusion are broad enough so a repository would have had to change its role completely to no longer be included.

### 6.2.3 Adding Records to the Registry

Entering metadata for a new record into the registry is a two-stage process. The first stage asks for the homepage URL for the repository. Clicking the ‘Next’ button performs some automated metadata extraction from the given URL and then presents the metadata entry form.

If the URL given by the submitting user is live and a normal web page the site’s title is extracted (the text contained in the HTML header title field). The domain is matched against a domain country list (with an additional mapping of `.edu` to United States), *e.g.* `ac.uk` is mapped to ‘United Kingdom’. International domain names are mapped into their ISO equivalent (*e.g.* `.org` is simply organisation), and rely on either the user or moderator to be corrected if the site is actually country-specific. The site is tested for the presence of GNU EPrints or DSpace OAI base URLs, respectively `/perl/oai2` or `/dspace-oai/request/`. If a response is forthcoming from the OAI base URL the software version is set (GNU EPrints or DSpace), and the OAI response parsed for the administrator’s email address (part of the OAI Identify verb).

The metadata entry form (Figure 6.1) provides the user with text entry boxes for all the ROAR metadata fields (Table 6.1), except the timestamp and remote IP address that are stored automatically and transparently. The available options for country, software, type and full-text are from controlled vocabularies, so are dependent on the user or ROAR moderator to provide appropriate choices and to fit ambiguous entries into the most appropriate category. All other fields can contain any text (although obviously the repository URL and OAI base URL only make sense as web URLs). The name, country, software, OAI base URL and email are automatically extracted where possible.

The screenshot shows a metadata entry form on the left and a preview of the resulting ePrints page on the right.

**Form Fields:**

- URL:**
- Name:**
- Country (if applicable):**
- Software:**
- OAI BaseURL:**
- Type:**
- Public Full-Text:**
- Status:**
- Comment/Description:**
- Admin's Email:**
- Submission Host:**
- Add Archive:**

**Preview (Right):**

The preview shows a page titled "Faculty of Technology ePrints Service". It includes a navigation menu with links: Home, About, Browse, Search, User Area, and Help. The main content area says "Welcome to the Faculty of Technology ePrints Service" and mentions "running GNU EPrints 2.3.6".

**Figure 6.1:** ROAR metadata entry form with embedded preview.

Once the metadata entry is finished the entry is stored in buffer of new records. The records awaiting acceptance are not made public, although attempting to enter the same record twice will result in an error. In order for a new record to go into the live listing it must first be checked and possibly corrected by a moderator.

A moderator logs in by clicking the 'Login' link and entering a username and password. When a moderator logs in the 'Register an Archive' link changes to 'Add an Archive', and they have an additional option of accepting a record from the new records buffer. A drop-down list of new sites is given along with 'Next' and 'Safe' links. The 'Safe' option suppresses the site preview (as some sites cause a redirect, or launch new windows). Selecting a site and clicking 'Next' shows the standard metadata entry form (Figure 6.1). When accepted the record goes into the live listing, hence is browseable and searchable straight away. If rejected (by changing the 'status' to 'dead'), the record is simply deleted from the system (this tends to apply only to spammers). Sites that go into the 'Dead sites listing' must first go into the live registry and then have their status changed to 'dead'.

## 6.2.4 Maintaining the Registry

Only moderators can add or remove entries from the registry (by changing the status of a record from 'new' to 'live' or 'live' to 'dead'). Moderators can also directly edit entries *e.g.* to correct mistakes, most often as the result of a

**Table 6.1:** ROAR metadata fields

|             |   |
|-------------|---|
| URL         | The home page of the repository   |
| Name        | The name of the repository <i>e.g.</i> Aberdeen University Research Archive: AURA   |
| Country     | The country the repository is based in  |
| Software    | The repository software used <i>e.g.</i> GNU EPrints  |
| OAI-PMH URL | The base URL of the OAI-PMH interface   |
| Type        | What defines the ‘collection’ <i>e.g.</i> theses, e-journal, institutional, disciplinary                                  |
| Full-text   | What proportion of the content has a ‘full-text’ associated (full-text meaning a paper and metadata versus only metadata) |
| Comment     | Any general notes about the repository  |
| Email       | The administrator’s email address   |

repository administrator’s contacting me by email. I also occasionally go through the registry to remove incorrect OAI base URLs (*e.g.* where the user has just copied their home page URL into the OAI base URL field), that can cause problems when updating Celestial’s records (entries in ROAR aren’t automatically registered in Celestial, instead they are periodically imported from ROAR’s `ListFriends` interface<sup>8</sup>).

As an experimental feature users can register a username and password that are then associated with entries they register. They can then update their own entries once accepted (*i.e.* set to ‘live’) by a moderator.

Whether it is an user or moderator editing an entry they use the standard metadata entry form, except the normal ‘Add Archive’ button is replaced by ‘Update’ (Figure 6.2). After making any corrections and clicking the Update Archive link the live entry is updated. The moderator is shown a summary of the new record and links that allow either returning to where they were or editing the next relevant entry.

The ROAR currently has no facility to automatically check whether repositories are alive, apart from the fact that if an repository disappears a thumbnail can no longer be generated (although often the result is a ‘404’ error message, rather than the server disappearing completely).

<sup>8</sup>`ListFriends` is a very simple XML listing of URLs <http://archives.eprints.org/index.php?action=listfriends>

Error connecting to sherpa.bl.uk: No route to host.

|   |   |
|---|---|
| <b>URL</b>                                    | <input type="text" value="http://sherpa.bl.uk/"/>             |
| <b>Name</b>                                   | <input type="text" value="British Library Research Archive"/> |
| <b>Country (if applicable)</b>                | <input type="text" value="United Kingdom"/>                   |
| <b>Software</b>                               | <input type="text" value="GNU EPrints"/>                      |
| <b>OAI BaseURL</b>                            | <input type="text" value="http://sherpa.bl.uk/perl/oai2"/>    |
| <b>Type</b>                                   | <input type="text" value="Research Cross-Institution"/>       |
| <b>Public Full-Text</b>                       | <input type="text" value="100%"/>                             |
| <b>Status</b>                                 | <input type="text" value="Live"/>                             |
| <b>Comment/Description</b>                    | <input type="text" value=""/>                                 |
| <b>Admin's Email</b>                          | <input type="text" value="mailto:support.sherpa@bl.uk"/>      |
| <b>Record Owned by</b>                        | <input type="text" value="tdb01r"/>                           |
| <b>Approved By</b>                            | <input type="text" value="tdb01r"/>                           |
| <b>Submission Host</b>                        | <input type="text" value="67.68.249.186"/>                    |
| <input type="button" value="Update Archive"/> |   |

Figure 6.2: Updating an entry in ROAR

## 6.2.5 Creating Web Page Thumbnails

Each entry in the ROAR has a small (150 by 120 pixel) ‘thumbnail’ of the repository’s home page. The purpose of the thumbnail is to allow the user to quickly compare repositories by visual inspection. This is particularly useful when comparing installations using the same repository software, as it is clear from the thumbnail how customised the software has been compared to the standard distributed version. (Many DSpace sites have only customised the generic DSpace logo to a local equivalent *e.g.* Figure 6.3, compared to more established repositories that have different layouts and added features *e.g.* Figure 6.4)



Figure 6.3: Minor customisations of DSpace





**Figure 6.4:** Major customisations of DSpace

When developing the ROAR it became apparent that no reliable tool was available for generating thumbnails of web pages, so a script for generating thumbnails was written from scratch. As the appearance of web pages is dependent on the tool used to view them the only practical solution to rendering and capturing an image of a web page is to use an existing web browser. As most web authors will only check whether their pages look correct using common browsers it is necessary to use a common browser to generate accurate previews of the linked repository.

The Mozilla web browser (Firefox) supports a scripting language called XUL that allows applications to be built using its graphical interface. This was used to automate the loading and screen-capture of repositories for the ROAR. The capture application consists of an application layout (described using XML) and javascript automation. The application layout is a featureless window (no titlebar, menus etc.) into which the repository's web site is loaded. When the application is run a javascript script is executed that starts a periodic check for errors, a timeout, adds an event handler to wait for the web page to load and lastly loads the requested URL. When the web page load event is completed the application calls the X window capture utility *xwd*. Optionally the application can repeatedly scroll and capture the window (if the web page is bigger than a single screen) to create a sequence of pictures from vertical sections of the web page. These can be used to create an animated thumbnail showing the entire page (although due to the large download size these were removed from ROAR).

In order for the repository thumbnail creation to work autonomously errors generated by Mozilla need to be handled automatically. These can consist of network errors (*e.g.* the repository is unreachable), secure HTTP certificate issues (*e.g.* certificate not signed by a recognised authority), or infinite redirection loops. Most DSpace installations by default use secure HTTP but don't have a

|                                  | Archives | In Celestial | Records | Mean | Median |
|----------------------------------|----------|--------------|---------|------|--------|
| • <a href="#">United States</a>  | 179      | 123          | 717882  | 5836 | 262    |
| • <a href="#">United Kingdom</a> | 69       | 57           | 104441  | 1832 | 229    |
| • <a href="#">Germany</a>        | 61       | 51           | 135462  | 2656 | 401    |
| • <a href="#">Brazil</a>         | 42       | 32           | 99014   | 3094 | 59     |
| • <a href="#">Canada</a>         | 32       | 30           | 28274   | 942  | 130    |
| • <a href="#">France</a>         | 29       | 26           | 126011  | 4847 | 409    |
| • <a href="#">Australia</a>      | 25       | 19           | 63532   | 3344 | 416    |
| • <a href="#">Sweden</a>         | 25       | 21           | 27863   | 1327 | 553    |
| • <a href="#">Italy</a>          | 22       | 19           | 12415   | 653  | 100    |
| • <a href="#">India</a>          | 19       | 13           | 8518    | 655  | 286    |

Figure 6.5: ROAR country-breakdown

valid certificate, resulting in warning messages. Unfortunately Mozilla doesn't provide an event model for handling errors, instead it shows a dialog box in the expectation that there is a user available to accept or cancel it. In the absence of a monitoring user the thumbnail application periodically checks for the presence of dialog boxes and accepts any errors to allow the web page loading to continue. If all else fails the application times out after 30 seconds, allowing the next repository to be processed.

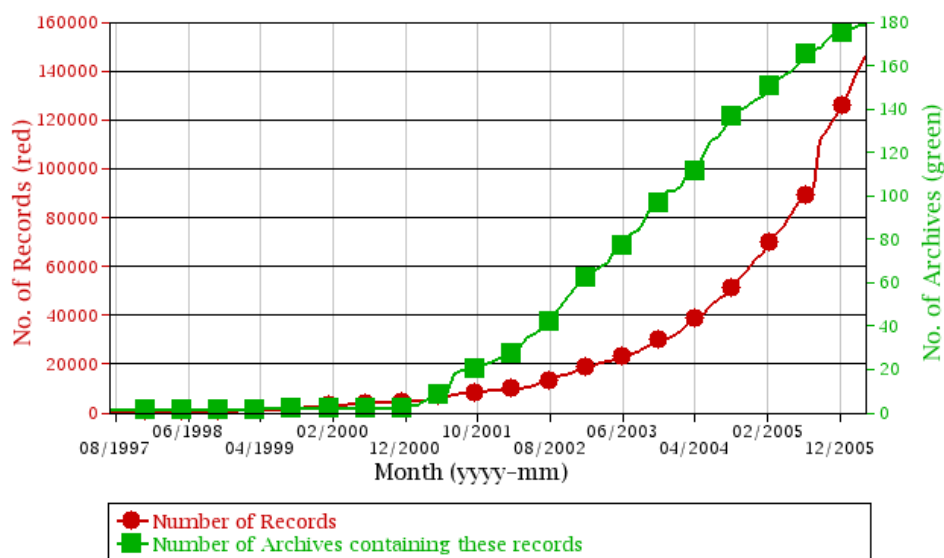
## 6.3 Analysis of Institutional Repositories

The ROAR provides a breakdown of institutional repositories by country, 'Archive Type' and institutional repository software under the 'Browse' page. The breakdown consists of the value (*e.g.* the country name), the number of repositories in that value, the number registered in Celestial and the number of OAI records found (Figure 6.5).

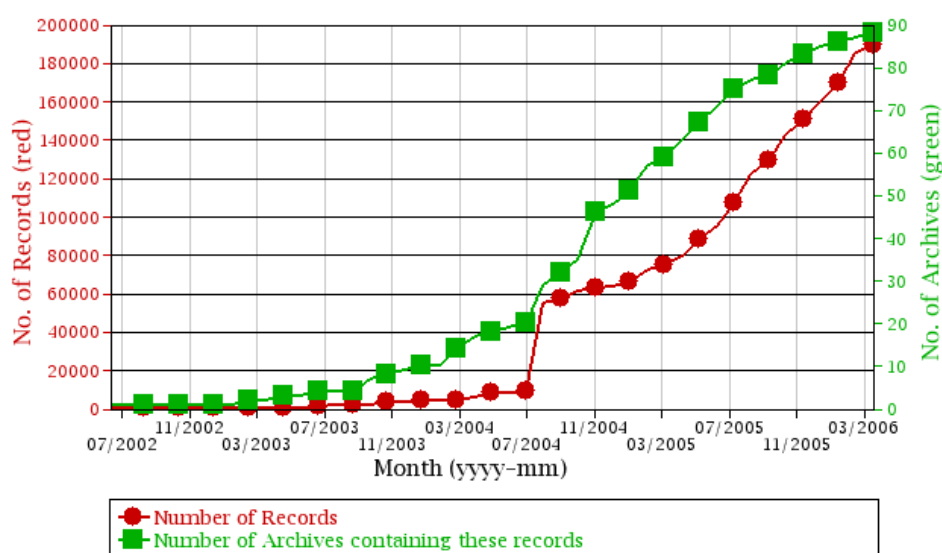
To analyse changes over time graphs can be generated for individual repositories or for any the available criteria (country, type or software, see *e.g.* Figure 6.6). Multiple criteria can be combined together to *e.g.* find all the Bepress-based repositories in the United States.

### 6.3.1 Software Installations

ROAR has thirteen predefined repository software types (see Table 6.2), accounting for most of the registered repositories. 186 repositories either use



**Figure 6.6:** Rate of growth of GNU EPrints-based repositories and contents



**Figure 6.7:** Rate of growth of DSpace-based repositories and contents

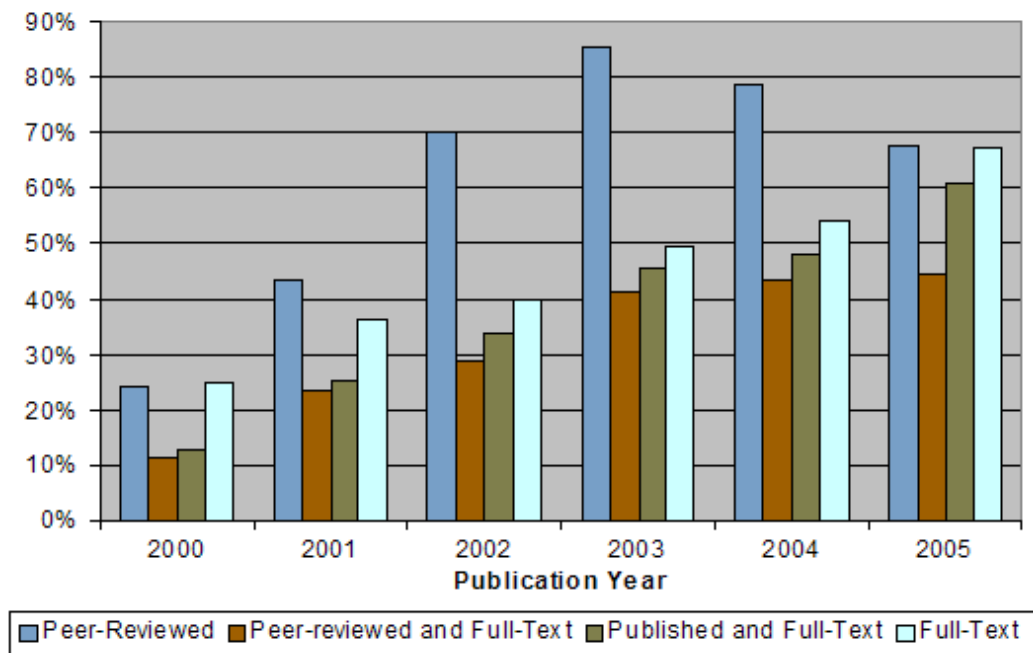
bespoke or unknown software types (many of these are databases rather than repositories of author self-archived content). With 200 installations GNU EPrints is the most widely installed repository software, with installations growing at a linear rate since 1997 (Figure 6.6). Similarly DSpace – the second most widely installed software – has seen a roughly linear increase in its installed base since its introduction in 2002 (Figure 6.7). Unfortunately some repositories either don't have an OAI-PMH interface, or an interface that isn't working, resulting in the summary graphs tracking the contents of fewer repositories than are registered in ROAR.

**Table 6.2:** Software types configured in ROAR

| Software                        | Entries | Celestial | Records | Mean  | Median |
|---------------------------------|---------|-----------|---------|-------|--------|
| GNU EPrints                     | 200     | 181       | 146529  | 810   | 164    |
| DSpace                          | 135     | 89        | 189384  | 2128  | 428    |
| Bepress                         | 43      | 25        | 58743   | 2350  | 510    |
| ETD-db                          | 22      | 18        | 264008  | 14667 | 1316   |
| OPUS (Open Publications System) | 21      | 18        | 5073    | 282   | 79     |
| DiVA                            | 14      | 13        | 9284    | 714   | 398    |
| CDSWare                         | 9       | 6         | 135494  | 22582 | 13597  |
| ARNO                            | 5       | 5         | 216214  | 43243 | 30819  |
| DoKS                            | 3       | 3         | 2172    | 724   | 226    |
| HAL                             | 3       | 3         | 55197   | 18399 | 1092   |
| Fedora                          | 3       | 3         | 1223    | 408   | 167    |
| EDOC                            | 2       | 2         | 40360   | 20180 | 20180  |
| MyCoRe                          | 1       | 1         | 1954    | 1954  | 1954   |
| Other                           | 186     | 128       | 2742117 | 21423 | 598    |

A higher proportion of GNU EPrints sites (90%) than DSpace (66%) have OAI-PMH interfaces registered with Celestial (Table 6.2). (GNU EPrints tends to have only one location for its OAI interface (`/perl/oai2`) compared to several variations for DSpace, so the difference may be due to not being able to locate the appropriate OAI-PMH URL for the DSpace repositories.) While DSpace accounts for fewer repositories it actually accounts for more OAI records because DSpace repositories, on average, contain two and a half times as many OAI records as GNU EPrints repositories. When examining the average number of records per repository, ROAR ignores repositories that Celestial was unable to harvest.

With 9400 records the largest GNU EPrints repository is the University of Southampton: Department of Electronics and Computer Science repository



**Figure 6.8:** ECS EPrints contents, based on its ‘Advanced Search’

(ECS). Of those 9400 records 2135 are in the OAI-PMH set Full text attached, meaning the metadata record has at least one digital document available online. 210 records have a full-text attached but are not made publicly available. Using ECS’s Advanced Search tool a more in depth look at the type of records contained is easily achieved (based on the author-supplied metadata).

The metadata definitions used in the ECS repository are subtly different to my definitions (e-prints that are, or a destined, for peer-reviewed publication). 1391 records were flagged by ECS as peer-reviewed and full-text, of which 536 were papers, 773 conference items, and 82 other records (book sections etc.). However, the peer-reviewed flag in ECS refers to the version of the full-text attached, so may exclude those e-prints destined for peer-review (pre-prints). The publication status of an e-print (unpublished, submitted, in press, or published) may provide a better clue for the type of record, but does not necessarily imply that the publication is a peer-reviewed, research journal. **Figure 6.8** shows for the latest records (2005 – 364 records in total) between 40 and 60 percent of records contain full-texts of published and/or peer-reviewed papers. This represents a considerable increase in author self-archiving (*e.g.* in 2000 only around 20 percent of any records had a ‘Full-Text’ attached). Part of this increase may be due to

the ECS department mandating full-text deposit by faculty.

The largest DSpace repository is The Australian National University (ANU) with 42858 records. Unfortunately ANU doesn't provide a simple publicly accessible means to determine the type and number of digital items (*vs.* metadata-only records). Using the OAI Registry at UIUC it is possible to create a break down of the number of records in each OAI set exported by ANU. 39364 records appear to be from a picture repository (contain JPEG images only). The largest non-picture category with 436 records is Business and Economics – Economics. Examining the first records returned in this set their type is either unspecified or 'techreport (published by ANU)'. It is likely that – as with Southampton/ECS – only a fraction of the total records in ANU are peer-reviewed research papers.

Southampton ECS – the home-site of GNU Eprints – provides all the tools that EPrints has out of the box, in particular, search options to retrieve only peer-reviewed research papers by restricting search to only published, peer-reviewed full-text items. The home-site of DSpace (DSpace at MIT) provides essentially the same functionality as ANU *i.e.* no facility is provided by the search or OAI interfaces to restrict results to peer-reviewed or published records, or to only those records that have a publicly accessible full-text. While on the face of it DSpace repositories have been more successful at attracting records, the limited ability to interrogate those records makes it difficult to determine whether or not DSpace repositories contain more publicly accessible, peer-reviewed research literature *e.g.* DSpace at Cambridge<sup>9</sup> contains 150,000 objects of which 147,000 are machine-generated chemical molecular structures.

The ability to distinguish full-text, peer-reviewed literature from the general institutional research output (databases, images, and other media) is important not only for assessing author's support for publicly self-archiving their research papers (hence maximising usage and impact), but to end-user services – such as Citebase Search and OAIster - to allow users to restrict their queries to just the peer-reviewed literature.

---

<sup>9</sup>DSpace at Cambridge <http://www.dspace.cam.ac.uk/>

**Table 6.3:** Top 11 countries ranked by number of repositories

| Country        | Entries | Celestial | Records | Mean  | Median |
|----------------|---------|-----------|---------|-------|--------|
| United States  | 179     | 123       | 717882  | 5836  | 262    |
| United Kingdom | 69      | 57        | 104441  | 1832  | 229    |
| Germany        | 61      | 51        | 135462  | 2656  | 401    |
| Brazil         | 42      | 32        | 102717  | 3210  | 59     |
| Canada         | 32      | 30        | 28274   | 942   | 130    |
| France         | 29      | 26        | 126011  | 4847  | 409    |
| Australia      | 25      | 19        | 63532   | 3344  | 416    |
| Sweden         | 25      | 21        | 27863   | 1327  | 553    |
| Italy          | 22      | 19        | 12415   | 653   | 100    |
| India          | 19      | 13        | 8518    | 655   | 286    |
| Netherlands    | 18      | 15        | 379508  | 25301 | 5359   |

### 6.3.2 Countries

ROAR has entries for institutional repositories in 41 different countries. A repository is determined to belong to a country if its content is submitted by users primarily in that country *e.g.* an institutional repository is classified as belonging to the country in which the institution is based.

**Table 6.3** provides a break down of the number of repositories and OAI records for the top 11 countries by most repositories registered. The total and average number of records (as found by Celestial) are shown for each country. The United States has by far the largest number of entries and records, although the Netherlands (at rank 11) has the largest number of records *per-repository*.

The US (103), UK (45), Germany (27), Australia (19), Canada (16), Sweden (16), Netherlands (13) and Italy (11) have the largest number of institutional repositories but, relative to their populations, Sweden, Netherlands, Australia, New Zealand, Norway and United Kingdom have the highest proportion of institutional repositories. The Netherlands has 100% availability of research repositories in its higher-education research institutions ([Van der Kuil and Feijen, 2004](#)).

**Table 6.4** provides a comparison between countries normalised by their population<sup>10</sup>. The table includes only those institutional repositories flagged as research institutional or departmental and that could be harvested by Celestial.

<sup>10</sup>Source US Census Bureau <http://www.census.gov/ipc/www/idbsprd.html>

**Table 6.4:** Top 11 countries ranked by number of institutional records, normalised by population

| Country        | Population | Celestial | Records |
|----------------|------------|-----------|---------|
| Netherlands    | 16491461   | 13        | 375899  |
| Switzerland    | 7523934    | 4         | 92690   |
| Australia      | 20264082   | 15        | 53713   |
| Finland        | 5231372    | 2         | 13393   |
| Sweden         | 9016596    | 15        | 20451   |
| Greece         | 10688058   | 1         | 23855   |
| Canada         | 33098932   | 16        | 23252   |
| United Kingdom | 60609153   | 37        | 34993   |
| United States  | 298444215  | 72        | 125628  |
| Portugal       | 10605870   | 1         | 4377    |
| Norway         | 4610820    | 3         | 1839    |

The total records from those institutional repositories is then divided by the country's population. Switzerland comes particular high due to the CERN Document Server (CERN), that contains 76078 records. The Netherlands has put a lot of effort into promoting institutional repositories, notably the DARE project ([Van de Vaart, 2004](#); [Van der Kuil and Feijen, 2004](#)), which seems to have paid off.

## 6.4 Peer-reviewed Full-Text Detection

A limitation on the analysis of institutional repositories is determining whether the publicly available records (via an OAI-PMH interface) have an associated publicly accessible full-text or other digital item. There is no OAI requirement or commonly implemented standard for describing the location of the resource described by an OAI record. Where a mechanism could be used to locate the full-text item an informative analysis would still need to determine the difference between research papers (published, peer-reviewed etc.) and all the other potential types of items (pictures, abstracts, multimedia content, databases, etc.).

As mentioned previously many GNU EPrints installations include an OAI set that contains records with a publicly accessible full-text item. For repositories without an equivalent set a web crawler could be used to check for the existence of a full-text using rules customised to each repository software. For DSpace each abstract page contains 'View/Open' links to the associated digital resources. A



web crawler could retrieve each DSpace abstract page, locate these links, and check the linked object's type (*e.g.* a PDF document might be defined as being a 'research paper', whereas a JPEG image might not).

Determining more accurately the difference between any digital file and a research paper requires looking inside the document. An initial check is simply whether the file is a text-capable format (PDF, Microsoft Word, Latex, *etc.*) *vs.* image or stream based media (JPEG, MP3, *etc.*). To determine the difference between any text document and research papers the end of the document can be searched for a bibliography, which journal and conference papers will almost always contain. Searching for a bibliography depends on being able to convert the document to plain text (or having an API capable of reading the document directly) and having rules capable of determining whether a block of text is a bibliography.

This kind of web crawling and research paper detection is already performed by the Citeseer and Google Scholar services, but they can't be restricted to institutional repositories and peer-reviewed research literature, nor do they provide tools to measure the number of publicly accessible peer-reviewed full-texts.

### 6.4.1 Full-text Detection Implementation

I have implemented a simple script to analyse the number of full-text items. The script combines an OAI-PMH harvester with some simple heuristics to identify full-text and research paper items.

In order to make testing for full-texts across many repositories quicker only a sample of records is taken from each repository. An initial request is made for all the record identifiers in the repository and from that set 1000 are randomly chosen. If the repository contains less than 1000 records then all its records are tested.

Typically the Dublin Core metadata harvested from an repository contains the URL of the abstract (or 'jump-off') page and sometimes includes direct links to the electronic full-text. Different repositories have used a number of Dublin Core fields to contain abstract and full-text links, even though each Dublin Core field is semantically different. The Dublin Core fields identifier, source, relation and

format are checked – identifier should be used to contain the full-text link (as it’s the URL of the resource), but source and relation have also been used by repositories that use identifier to link to the abstract page, and GNU EPrints has used format (which should only contain the resource file format).

Often access to journals is granted based on the user’s domain and, because the script is run from within the University domain, it could incorrectly identify a journal article as open access when in fact it is only available by subscription. Hence full-text URLs that are not on the same server as the repository’s OAI interface are ignored, to avoid counting the journal version of the paper. This caused zero matches for some repositories *e.g.* the CERN Document Server, where the OAI interface is hosted on `cdsweb.cern.ch` but links to `preprints.cern.ch` for full-texts.

The mime-type (the file format) supplied by the server was used to determine whether a document was in Microsoft Word (msword) or PDF (pdf) format. The total number of Word and PDF files provides an estimate of the number of research papers, as the (anecdotally) most widely used formats in IRs. In addition each format that could be reliably converted to plain text (plain-text, HTML, PDF and Word) was parsed for something that looks like a bibliography. This is based on looking for lines containing 4-digit numbers starting with ‘19’ or ‘20’ (*i.e.* years) up to 300 characters apart. If at least 5 matching lines are found the document is determined to be a ‘research paper’.

### 6.4.2 Full-text Results

The algorithm used to detect full-text research papers is not intended to be rigorous, but rather a quick method to get indicative figures for heterogeneous software and collections.

**Table 6.5:** Number of OAI records pointing to full-texts

|                           |        |
|---------------------------|--------|
| Total Texts Harvested     | 59536  |
| % PDF or Word             | 53.63% |
| % Containing Bibliography | 36.17% |

Of 220 repositories that were indexed in Celestial, 79 were found that contained at least one full-text or ‘research paper’ (the others either not being compatible

with the full-text detection script or genuinely containing no full-text resources). Of the full-texts downloaded (see [Table 6.5](#)) 54% were in Adobe PDF or Microsoft Word formats, giving an indication of the maximum percentage of records likely to contain a full-text research paper. Of the total texts harvested (plain-text, HTML, PDF or Word) only 36% appeared to contain a bibliography (subject to systematic errors in conversion to plain-text and incorrect detection).

## 6.5 Conclusion

The ROAR and other registry tools allow high-level analyses to be performed on research repositories. I have found that there has been a rapid growth in both the number of repositories and the number of records they contain, but for most repositories only around a third of their records correspond to full-text research papers. It remains to be seen how successful institutional repositories will be in getting more of the research literature from their authors.

To make deeper and more informative analyses possible, research repositories need to support more detailed tools and expose the record type *e.g.* to allow the selective harvesting of only those records with a freely accessible full-text attached and only those tagged as preprints or peer-reviewed postprints. An alternative approach would be to use an authoritative database of published material to test author self-archived and self-tagged (*i.e.* author-supplied metadata) records against, for instance using bibliographic data from publishers to match against e-prints to determine whether the e-print is a version of a publisher paper. Unfortunately few publishers give their bibliographic data away for free, or at least in a form easily used to analyse IR's content.

# Chapter 7

## Building an Open Access Citation Index

### 7.1 Introduction

The ROAR (previous chapter) and Celestial ([chapter 5](#)) provide high-level analyses of open access repositories. These analyses have been extended using citation and download data, based on an autonomous citation index I have built: Citebase Search.

Citebase Search is a live service serving about 10000 visitors per day<sup>1</sup>. It provides impact metrics and citation navigation for author self-archived papers deposited in several OAI-compliant repositories. These records can be searched and the results rank-ordered by per-paper or per-author citation or web download impact. Metadata and citation data are harvested autonomously, normalised, linked and stored in a structured MySQL database.

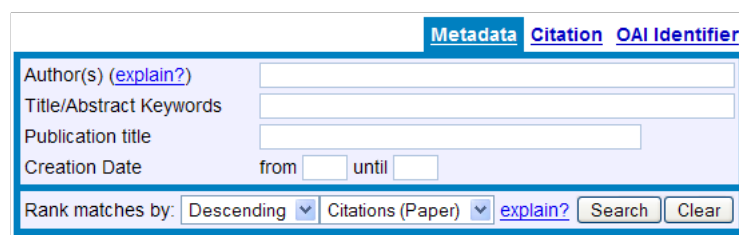
Citebase Search contains 400,000 full-text e-print records from which approximately 12.5 million references have been parsed, with 2.9 million references linked to the full-text (see [Figure 7.4](#)). 75% of the records contained in Citebase Search are harvested from arXiv (see [section 8.3](#)). The other records come from the institutional repositories at the University of Southampton,

---

<sup>1</sup>Based on usage analysis, see [section 7.7](#)

Biomed Central (an open access publisher), and several other small GNU e-prints based repositories.

Citebase Search harvests metadata for each paper using the OAI-PMH and augments this by extracting the references for each full-text paper. In addition, usage statistics are downloaded from the UK arXiv mirror and University of Southampton repositories. The association between document records and references is the basis for a classical citation database (similar to Web of Science or CiteSeer). Citebase Search can be thought of as a kind of “Google for the refereed literature”, as it ranks search results based on the number of references to papers or authors. Google combines a search relevance score with a page ranking algorithm calculated from the number of web links to a page. While currently Citebase Search only uses an absolute number of links, this could be expanded to use ‘hub-authority’ type algorithms by combining the search score with an authority weighting *e.g.* similar to the PageRank algorithm Google uses (Brin and Page, 1998).

The image shows a web interface for searching Citebase Search by Metadata. At the top, there are three tabs: 'Metadata' (selected), 'Citation', and 'OAI Identifier'. Below the tabs, there are four input fields: 'Author(s) (explain?)', 'Title/Abstract Keywords', 'Publication title', and 'Creation Date'. The 'Creation Date' field has 'from' and 'until' sub-fields. At the bottom, there is a 'Rank matches by:' section with two dropdown menus: 'Descending' and 'Citations (Paper)'. To the right of these are links for 'explain?', 'Search', and 'Clear'.

**Figure 7.1:** Searching Citebase Search by Metadata query.

I developed Citebase Search using data from the JISC/NSF Open Citation Project (see [subsection 3.3.3](#), page 38), which ended December 2002. I have continued to develop Citebase Search beyond the life of the OpCit project, and Citebase Search is now directly linked from arXiv (fulfilling OpCit’s goal of providing citation linking for arXiv). As part of the final OpCit report a user survey was conducted on Citebase Search (Hitchcock et al., 2003). This was used both to evaluate the outcomes of the project and to help guide the future direction of Citebase Search as an ongoing service. The report found that “Citebase can be used simply and reliably for resource discovery. It was shown tasks can be accomplished efficiently with Citebase Search regardless of the background of the user.”

Primarily a user-service, Citebase Search provides a web site that allows users to perform a metadata-search (query by title, author *etc.*), navigate the literature

using linked citations and citation analysis, and to retrieve linked full-texts in Adobe PDF format. Citebase Search also provides a machine interface to the citation data it builds through its own OAI-PMH interface *i.e.* other servers can download the same metadata as harvested from the source repositories, but augmented with citation links.

## 7.2 Citebase Search and the OAI-PMH

Citebase Search makes use of the OAI-PMH to harvest metadata from e-print repositories. A list of the repositories' OAI-PMH base URLs is stored, along with the date of the last harvest from each repository. Citebase Search requests any new or changed records from each repository since the last time Citebase Search successfully harvested. (The time used is the time at the start of the previous harvest, otherwise records added during a harvest may be missed in the next update.) An overlap of one day is added to be safe (the same records may be harvested more than once, however that is preferable to missing records).

The OAI-PMH API library<sup>2</sup>, developed as part of Citebase Search (also used in Celestial), provides an abstract interface to OAI-compliant repositories. This hides the complexity of OAI flow-control and error-handling. Using the API for harvesting consists of creating a repository object that contains the base URL and then calling the OAI-PMH commands on that object (*e.g.* ListRecords to list all metadata records). The library returns a list of objects that contain the metadata, or an error code and message. Only if the harvest completes successfully is the harvest time updated. In the event of an error the harvest is started again at the next scheduled update.

The arXiv and GNU EPrints-based repositories are harvested via Celestial (chapter 5), to take advantage of its caching and error handling facilities. Biomed Central<sup>3</sup> is harvested directly; BMC exposes the full-text through its OAI-PMH interface, which is too large for Celestial to store, so Citebase harvests directly from BMC's OAI interface to extract citations from the full-texts.

<sup>2</sup>HTTP::OAI::Harvester <http://search.cpan.org/~timbrody/HTTP-OAI-3.16/>

<sup>3</sup>Biomed Central <http://www.biomedcentral.com/>

## 7.3 Reference Parsing

Citebase Search’s reference parsing handles structured and unstructured documents. I use the term ‘structured’ to describe a document where there is some semantic description of its content *e.g.* most TeX documents have references that are labeled in a (TeX-command) ‘bibliography’ section. ‘Unstructured’ documents – such as HTML – may have markup around a title (*e.g.* a HTML heading, <H> tag), but is typically used by authors as a stylistic description (big font, bold) rather than saying ‘this is the title of the document’, leading to ambiguity. The supported ‘structured’ formats are TeX or XML, available respectively from arXiv and Biomed Central. Unstructured formats (PDF, plain text, HTML) are processed using heuristics to retroactively identify the document structure.

### 7.3.1 TeX

The first method used by Citebase Search (developed by the Open Citation Project) to extract references from documents is by parsing the ‘bibitem’ entries from TeX-format documents. These bibitem mark-ups (literally `\bibitem`) enclose a free-text string containing a citation, and possibly simple macros (*e.g.* to simplify writing journal titles). Depending on the author’s citation style it may contain a reference to more than one real-world paper *e.g.* see reference 2 in [Figure 7.2](#).

- [4] I. Antoniadis and M. Quirós, *Phys. Lett.* **B392** (1997) 61.

[5] I. Antoniadis, *Phys. Lett.* **B246** (1990) 377; I. Antoniadis, C. Muñoz and M. Quirós, *Nucl. Phys.* **B397** (1993) 515.

**Figure 7.2:** Compound reference.

The Citebase Search parsing code adds another tag around each Bibitem (`xxxOpcit`) then runs TeX over the document to generate an intermediary ‘DVI’ format document. The DVI file contains the xxxOpcit tags and the output text of the document. The unstructured text of each reference is extracted from the DVI file.

Processing the TeX source is necessary to expand any macros that the author may have included within the references (which would otherwise require writing a custom TeX processor to handle). By processing the TeX directly the entire reference section is captured and that few character errors occur (because the output is always in ASCII encoding). The vast majority of references are also extracted one-citation per ‘line’, which avoids identifying and splitting compound references. The difficulty with parsing TeX is the multitude of TeX styles (macros *etc.*) that authors can use, leading either to problems executing TeX at all or spotting the type of reference mark-up an author has used (if the author hasn’t used `bibitem`).

### 7.3.2 XML

Documents from Biomed Central are downloaded from their OAI interface in XML format. These documents are fully ‘marked-up’, hence the references already contain machine-parseable metadata (author names, bibliographic data *etc.*) – [Figure 7.3](#) shows an example. Each reference is contained in a ‘bibl’ structure that is extracted and the component data parsed and stored in the database.

```
<bibl xmlns="http://www.biomedcentral.com/oai/2.0/XML/schemas/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.biomedcentral.com/oai/2.0/XML/schemas/ http://www.biomedcentral.com/oai/2.0/XML/schemas/bibl.xsd" ui="rr68" >
<title>
<p>Airway obstruction in asthma: does the response to a deep inspiration matter?
</p>
</title>
<aug>
<au id="A1" ca="yes">
<snm>Fredberg</snm>
<mi>J</mi>
<fnm>Jeffrey</fnm>
<insr iid="I1"/>
<email>jfredber@hsph.harvard.edu</email>
</au>
```

**Figure 7.3:** XML format reference.



### 7.3.3 Unstructured Documents

Unstructured documents contain no mark-up within the document to indicate what the text means (*e.g.* is ‘Smith’ an author’s name or a profession?).

Postscript, Adobe PDF and (most) HTML documents only contain unstructured text (albeit with varying degrees of typographical structure). To parse the references from these types of documents Citebase Search converts them to plain text, then uses heuristics to find the references.

If the text is in two columns it is de-columnised by finding the modal distance from the left margin where a large space occurs (a check is made by finding if this distance is approximately 50% of the mean line length), and splitting each line around that point. De-columnizing takes account of differing whitespace to the left of the page, where the document has different layouts on alternating pages (often the case for media destined for print, where the book-margin alternates from right to left on double-sided printing).

The first step to extracting the references from the plain text is to locate the reference section. Citebase Search does this by locating a title containing the word ‘reference’, ‘bibliography’ or ‘notes’. A title is a line of text surrounded by empty-lines, either preceded by a number (*e.g.* ‘20. References’) or capitalised (*e.g.* ‘REFERENCES’). If such a title could not be found, a section containing sequentially numbered lines is looked for.

A number of rules are compared to the text below the reference’s title to separate out individual references and to find the end of the reference section (figures and acknowledgements may be at the end of the document). The simplest reference style to parse is numbered, either ‘1.’ or ‘[1]’. If the references are unnumbered they may be separated by whitespace, or by lines containing the publication year (*e.g.* ‘John Smith and Jane Doe ... (1993) ...’).

The reference parsing facility of Citebase Search is a set of cases evolved over time that provide a good coverage of the literature Citebase Search is required to process. The documents that can not be parsed may not be convertible or convert badly (*e.g.* postscript conversion often results in garbage). Once in plain-text the ability to parse the references is dependent on the format and style used by the author. Every author will use slightly different styles/conventions – autonomous reference parsing works on a ‘best case’ rather than complete basis.

### 7.3.4 Citation linking

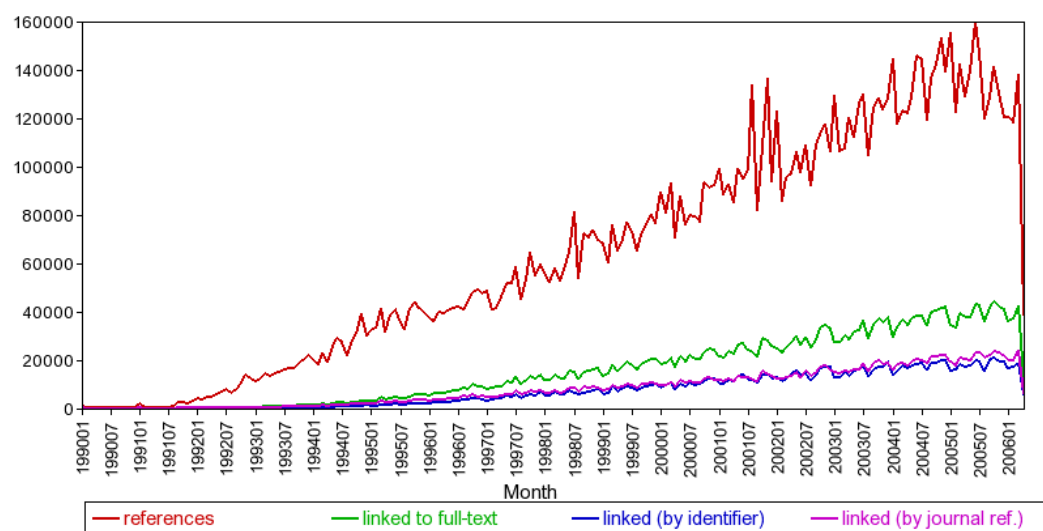
With successful extraction of the references from the full text, Citebase Search can parse individual references into their components: authors, title, year of publication, publication title (journal), identifiers, *etc.* The citation components can then be used to locate the full-text record of the cited paper. This creates links between citing and cited papers, which, for a collection of papers, creates a citation database.

Citations are intended as unambiguous references to a real-world thing. They do this through providing a meta-description of the cited thing. In most cases this is a bibliographic reference to a journal paper *e.g.* consisting of author, year of publication, journal and issue and page reference. As a database it is convenient to store numbers, as it makes it quick to locate matches. A bibliographic reference can be reduced to 3 numbers (year, issue and page number) with an author and/or journal name to avoid false-positive duplicate matches. If Citebase Search is unable to find a match using the bibliographic data it attempts to find a match by using the author and paper title (if given). In order to speed up the matching process Citebase Search normalises titles by removing very short words ('a', 'the', 'and' *etc.*) and only includes up to 5 words. This can result in a false match if an author has created two papers with near identical titles.

Citebase Search has two sets of papers that it knows about. The first are papers deposited by authors in e-print repositories, which are then harvested using OAI-PMH. The second set are those papers cited by the harvested papers, but are not available within the e-print repositories ('offline' papers). While Citebase Search may not be able to provide a user with the full-text of an offline paper Citebase Search can still track the citations to these offline papers, and provide a citation impact score and other analyses. These offline papers are also included in each author's citations, providing a more accurate picture of their citation impact and publication list (obviously there is no web impact score for an paper that is offline!).

**Figure 7.4** shows the number of references parsed from full-text papers and the number of those references that have been successfully linked to the cited paper. The number of papers linked using an author-provided arXiv identifier and those linked using a journal reference are roughly equal, and together allow 30% of references to be linked. The percentage of references that can be linked has

increased slightly over the years due to the number of citable papers increasing (the larger the collection of papers in Citebase Search, the more likely it is a cited paper is available, as well as an increasing historical backlog of papers).



**Figure 7.4:** Total references parsed and linked in Citebase Search, per month.

## 7.4 Citebase Search's Web Interface

Citebase Search is a combination of a basic search engine and citation database, enabling search results to be ranked by citation metrics and to navigate the literature using citation links. The initial search screen shown in [Figure 7.1](#) (page 94) provides three search choices: query by metadata (authors, search term, publication title and creation date *e.g.* [Figure 7.5](#)), query by citation (a bibliographic reference) and a number of queries by record (OAI) identifier ([Figure 7.12](#), page 104, shows querying for co-citing papers using the OAI Identifier query).

Clicking the title-link in a result set shows an 'abstract' page for the selected record. The top of this page contains the author-supplied metadata for the record (title, authors, abstract *etc.*), as shown in [Figure 7.6](#). Links to the search engine for each author allow searching for other papers by the same named author (Citebase Search does not differentiate between authors with the same name). As a citation index for open access literature Citebase Search can provide direct links to the full-text and – in the case of arXiv – a locally cached PDF. The Linked

The screenshot shows a search interface with tabs for 'Metadata', 'Citation', and 'OAI Identifier'. The 'Metadata' tab is active. Search criteria include: Author(s) 'Witten, E', Title/Abstract Keywords 'string theory', and Creation Date range. The results show two entries: 'Anti De Sitter Space And Holography' (2691 citations) and 'String Theory and Noncommutative Geometry' (1699 citations). Each entry includes a brief abstract snippet and a 'CrossRef' link.

Figure 7.5: Results from a metadata search for “Witten, E” and “String Theory”.

PDF link provides an experimental format where a CGI server inserts overlay links for references into the full-text PDF.

The screenshot shows an abstract page for the paper 'New Dimensions at a Millimeter to a Fermi and Superstrings at a TeV' by Antoniadis, I.; Arkani-Hamed, N.; Dimopoulos, S.; Dvali, G. The abstract text is highlighted in green. A jagged line is drawn across the text. Below the abstract, it says 'Full-text available from: Phys.Lett. B436 (1998) 257-263' and provides a URL: 'http://arxiv.org/abs/hep-ph/9804398'.

Figure 7.6: Citebase Search abstract page (jagged line is abbreviation).

A summary of the citations to and downloads of the paper is presented in tabular and graphical forms (in Figure 7.7). The graph provides a record of the citations and downloads over time, shown cumulatively (originally the graph showed the raw data – sufficiently noisy to be described as a ‘richter scale’ by one user!). The table includes links to a listing of all the citing papers and an in depth listing of the downloads of the paper. The downloads page shows the number of downloads of the paper broken down by country (Figure 7.8) and a plot of individual downloads per month indicated by flags for the country the user accessed from (Figure 7.9).

Figure 7.10 shows an example where Citebase Search has successfully parsed the references from a paper. The references (as formatted by the author) are printed

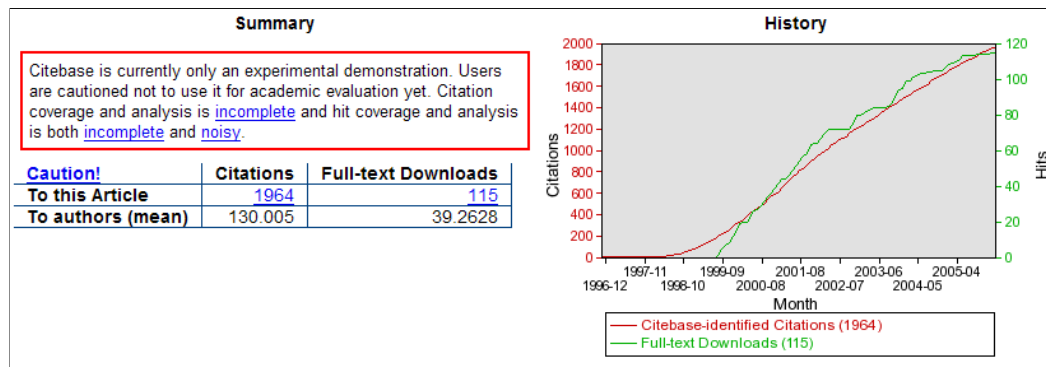


Figure 7.7: Download and citation summary table and figure.

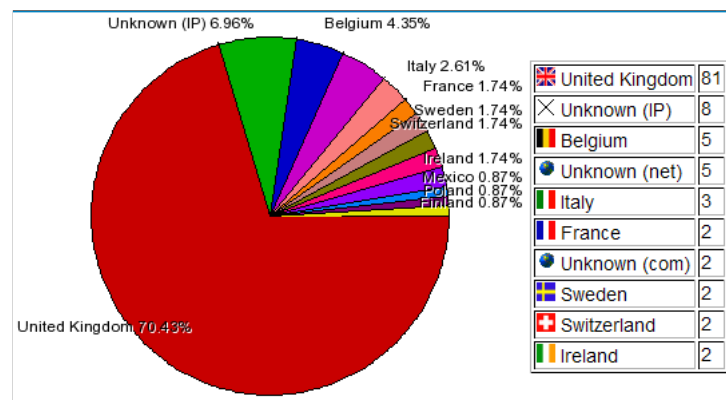


Figure 7.8: Downloads by country.

verbatim, along with links intended to making locating the cited paper easier. Where the cited paper has been found in Citebase Search an ‘eprint’ link is provided, taking the user directly to the cited paper. ‘eprint’ links are created either where the author has given an identifier for the cited paper (*e.g.* the top reference containing `hep-ph/9803315`) or where a matching cited paper contains the same bibliographic reference (the journal, volume, year *etc.* match). Heuristics are used for several journals to provide direct links to the journal version *e.g.* in this case ‘Phys. Rev. D.’ has been linked in reference 7. Where the cited paper couldn’t be located ‘G’ and ‘A’ links are provided that link respectively to Google Scholar and an author search in Citebase Search.

Below the list of references for the current paper two additional sections provide citation navigation to papers related to the current paper. Firstly – as a citation index the primary function of Citebase Search is to provide links to papers citing the current paper – the top 5 most cited papers citing the paper are shown, along with a link to search for more. Secondly the top 5 most co-cited papers are



**Figure 7.9:** Downloads over time.

|                        |  |
|------------------------|--|
| <a href="#">eprint</a> | [1] N. Arkani-Hamed, S. Dimopoulos and G. Dvali, hep-ph/9803315, to appear in Phys. Lett. B. |
| <a href="#">G/A</a>    | [2] S. Dimopoulos and H. Georgi, Nucl. Phys. B193 (1981) 150.                                |
| <a href="#">eprint</a> | [3] E. Witten, Nucl. Phys. B471 (1996) 135.  |
| <a href="#">eprint</a> | [4] I. Antoniadis and M. Quirós, Phys. Lett. B392 (1997) 61.                                 |
| <a href="#">G/A</a>    | [5] I. Antoniadis, Phys. Lett. B246 (1990) 377   |
| <a href="#">eprint</a> | I. Antoniadis, C. Munoz and M. Quirós, Nucl. Phys. B397 (1993) 515.                          |
| <a href="#">G/A</a>    | [6] C. Bachas, 1995, private communication.  |
| <a href="#">eprint</a> | [7] J. D. Lykken, <a href="#">Phys. Rev. D</a> 54 (1996) 3693.                               |
| <a href="#">eprint</a> | [8] P. Horava and E. Witten, Nucl. Phys. B460 (1996) 506 and B475 (1996) 94.                 |

**Figure 7.10:** Citebase Search reference list.

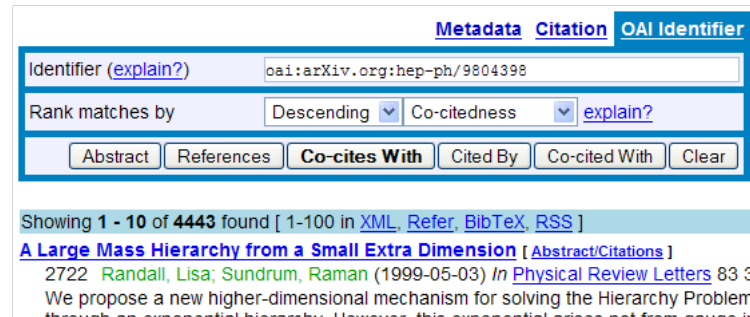
shown, again with a link to search for more. Both of the ‘top 5’ listings show a summary extract for each linked-to paper (similarly to the search results listing). [Figure 7.11](#) shows these summary blocks, which include the usual metadata (title, authors, abstract), as well as the total citations and total number of co-citations respectively for the two top 5 listings. Citing papers are useful to show as they provide a path to newer research on the same subject, similarly strongly co-cited papers are likely to be related, but may not directly cite (or be cited by) the current paper.

|  |   |
|--|---|
| <b><a href="#">A Large Mass Hierarchy from a Small Extra Dimension</a> [ <a href="#">Abstract/Citations</a> ]</b>  |   |
| 2722   | <a href="#">Randall, Lisa; Sundrum, Raman</a> (1999-05-03) <i>In</i> <a href="#">Physical Review Letters</a> 83 3370 (1999) |
| We propose a new higher-dimensional mechanism for solving the Hierarchy Problem. The Weak scale arises through an exponential hierarchy. However, this exponential arises not from gauge interactions but from the geometry of the extra dimension.            |   |
| LaTeX  |   |
| <b><a href="#">An Alternative to Compactification</a> [ <a href="#">Abstract/Citations</a> ]</b>   |   |
| 2386   | <a href="#">Randall, Lisa; Sundrum, Raman</a> (1999-06-08) <i>In</i> <a href="#">Physical Review Letters</a> 83 4690 (1999) |
| Conventional wisdom states that Newton's force law implies only four non-compact dimensions. We demonstrate that a non-factorizable background geometry is possible. The specific example we study is a single 3-brane embedded in a higher-dimensional space. |   |
| pages  |   |

**Figure 7.11:** Metadata summary (taken from ‘Top 5 citing papers’).

Papers that are bibliographically coupled with the current paper can be searched for via the main search engine (results shown in [Figure 7.12](#)). Coupled papers are papers that share common references with the current paper, with the more references shared the more related the two papers. Whereas co-cited papers rely

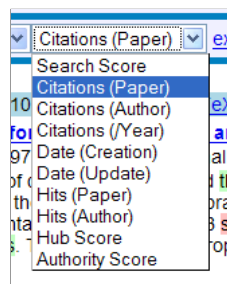
on there being citing papers (hence isn't available for new papers), coupled papers can be shown for any paper that contains linked references.



**Figure 7.12:** Bibliographically coupled papers search.

### 7.4.1 Ranking Search Results by Citation Impact

Citebase Search provides a number of different rank-orders for search results (Figure 7.13), found either by Boolean search query, or by the citation graph (*e.g.* papers that have cited the current paper). The live user service provides ten possible rankings: where a search has used a Boolean search query the results can be rank-ordered by the search score; date-ordered by either the accession date or last update of the paper; the number of citations: to the paper, authors' mean citations or average citations per year; the total downloads of the paper, or authors' mean downloads; and lastly, two experimental metrics provide ranking by hub or authority score (based on the HITS algorithm).



**Figure 7.13:** Citebase Search search result rankings.

The date rankings are taken directly from the dates exported by the e-print repository (see subsection 3.4.1 for discussion of the *Dublin Core* 'date' field) – the 'accession' being the earliest date and the 'last update' the latest.

The paper citation impact ranking is the total number of citations to an e-print. This ranking is skewed towards e-prints that have been around longer, because they have had longer to acquire citations. An author's impact is calculated as average citations to papers on which the author is named. The author impact score is the mean author impact of the named authors. A single-authored paper is hence equivalent to the average number of citations to that author's papers.

The download impact is calculated by the same method as the citation impact. Paper download impact is the total number of downloads of that paper. Author download impact is the mean of the named author download impacts. Download data is available only from the UK-based arXiv mirror, which reduces the total download score and may also skew the scores towards UK-centric papers.

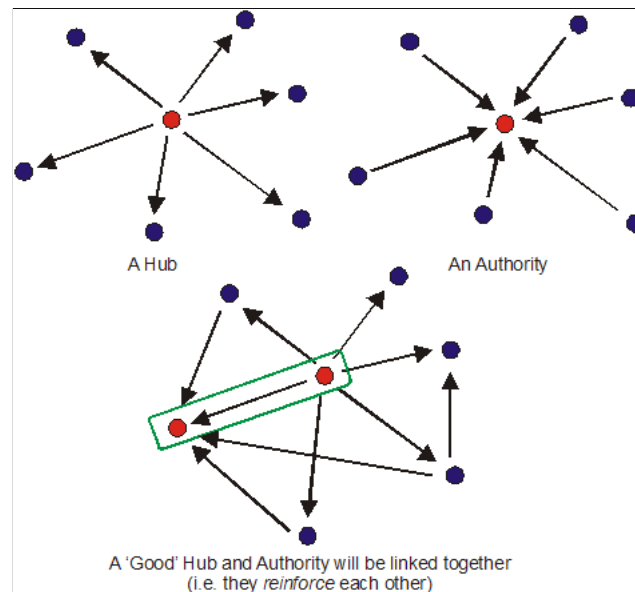
A hub is an index of information pages (*e.g.* a directory) and an authority a page that contains information on a particular topic. When using hub/authority for ranking web search results the Hub score isn't used – the assumption being that users aren't interested in finding directories. When applying hub/authority to the research literature authorities might be seminal papers on a topic and hubs may be review papers (*i.e.* directories of other papers). As review papers contain information (in the form of summaries) as well as just links to other papers, both the authority and hub scores can be used for rank-ordering in Citebase Search.

### 7.4.2 The Hub/Authority Algorithm

The hub/authority algorithm (Kleinberg, 1999) is a reinforcing algorithm that attempts to find the influence of a vertex within a directional graph (Figure 7.14). A good hub (a vertex with a high hub score) links to good authorities. A good authority (a vertex with a high authority score) is linked to by good hubs. Hub/authority was originally conceived to provide a way of finding good sources (or “authorities”) of information on the web – developed by Kleinberg as the HITS algorithm.

Hub/authority is calculated iteratively and converges. Each vertex (representing an e-print) contains four values: hub score, authority score, new hub score and new authority score. All values are initialised to 1. Each vertex is visited setting the new hub score equal to the sum of the authority scores of vertices that it links to, and the new authority score equal to the sum of the hub scores of vertices that





**Figure 7.14:** Hub-Authority algorithm (HITS).

link to it. The new Hub and Authority scores are then normalised so the sum of all hub scores is equal to 1, and the sum of all authority scores is equal to 1. Each vertex's hub and authority scores are set to the new values, and the next iteration started. The iteration is repeated 5 times, which finds a good approximation of the eventual (convergent) scores.

## 7.5 Citebase Search Database Structure

The Citebase Search database forms the basis for much of the subsequent work described in this thesis. The core database tables used are 'Links', 'Hits', 'Ranking' and 'Reference'.

The Links table provides the raw citation links between e-prints indexed by Citebase Search. This consists of the source citing identifier, target cited identifier, and a position field that allows the reference (in the Reference table) to be identified. Calculating the citation impact of an e-print involves counting the number of unique source identifiers for a given target identifier. Joining together multiple instances of the Links table allows co-citation and bibliographic coupling analyses to be achieved (analyses based on immediate neighbours in the network graph of citation links).

The Hits table contains the identifier of the downloaded e-print and the date/time the download occurred. The total downloads for an e-print can hence be counted or summarised over time (by-month, by-day *etc.*).

The Ranking table contains basic metrics for a given e-print and corresponds directly to the possible rank-orderings the Citebase Search web interface provides. Of particular interest is the accession date, which – when linked with the Links table – allows the time difference to be calculated between a citing and cited paper being deposited (its ‘citation latency’ – as will be discussed later).

The largest table in the Citebase Search database is the Reference table, that contains both the raw (authored) reference string, as well as its bibliographic components. The Reference table is rarely used directly due to its size (many millions of records), but is unavoidable when showing the number of references found over time, or the cited dates (as used in the analyses of literature obsolescence, discussed later).

All of these core tables are updated daily, as new records are harvested from source repositories. Some auxiliary tables are generated directly from these core tables to facilitate real-time analyses *e.g.* the citation latencies for all citation links. Otherwise, the data is read into analytical scripts that generate tables or graphs on-demand.

## 7.6 Citebase Search Analysis Tools

During the progress of my doctoral work I have created several tools to analyse Citebase Search’s database. These are predominantly web (CGI) scripts that use parameters supplied by the user to filter and process data read from the database, generating either a graph or table as output. The ‘statistics’ tool<sup>4</sup> (see [Figure 7.15](#)) provides fifteen different analyses, some of which have been used in other sections. The **references** analysis (the total number of references extracted and linked per month) was used earlier in this chapter, see [Figure 7.4](#). The analyses supported range from the trivial (number of new e-prints per month) to the quite complex (comparison of download and citation obsolescence). Each analysis is briefly described in the following subsections.

---

<sup>4</sup>Citebase statistics tool <http://www.citebase.org/cgi-bin/analysis/statistics>

```

Statistics Generator

This script generates a number of graphs/tables that provide basic
analysis of the data in Citebase.

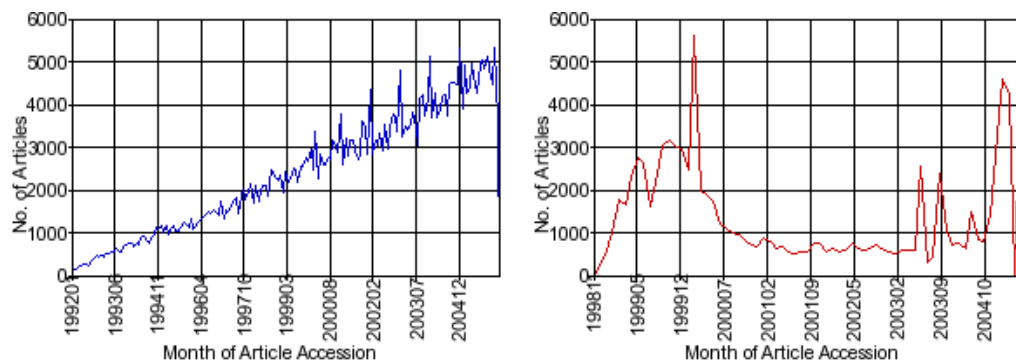
Usage:
    statistics?type=...&width=800&height=600&format=graph
Where:
    type = article frequency
           article frequency ads, citation frequency
           citation latency, citation latency per area
           citation latency per year, citation zipf
           cited age, explain, hits frequency
           hits latency, hits latency normalised
           hitsbydomain, hitslatencybyquartile
           papers per field, reference latency
           references, test graph
    format = graph|table

```

**Figure 7.15:** The statistics script provides links to all available analyses.

### 7.6.1 Paper Frequency: `article_frequency`, `article_frequency_ads`, `papers_per_field`

`article_frequency` shows the number of papers added per month based on the earliest date associated with the record – its accession. The `article_frequency_ads` generates the same data but for the data set from the NASA Astrophysics Data System<sup>5</sup> (ADS). Both analyses are shown in **Figure 7.16**. `papers_per_field` shows the number of new papers per annum for given arXiv subject areas *e.g.* see **Figure 8.2**, page 120.

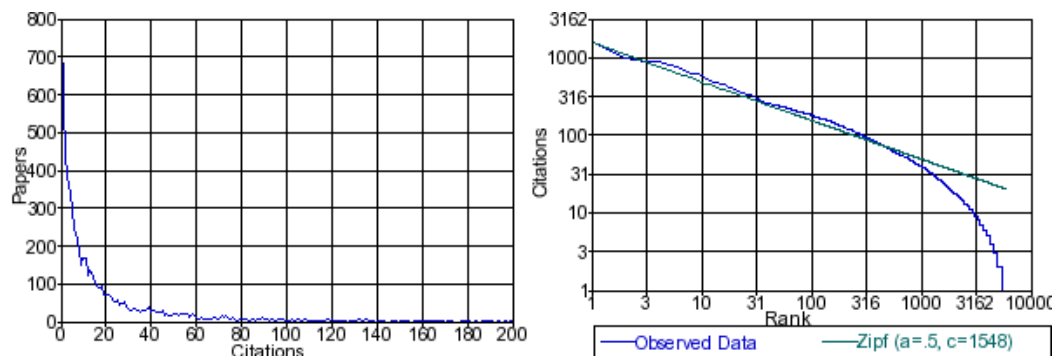


**Figure 7.16:** Papers frequency in arXiv (left) and the ADS data set (right)

<sup>5</sup>The NASA Astrophysics Data System <http://adswww.harvard.edu/>

### 7.6.2 Citation Histogram: `citation_frequency`, `citation_zipf`

**Figure 7.17** shows a histogram of papers by citation impact (citations per paper) – and a histogram of papers rank-ordered and plotted on logarithmic axis (a Zipfian plot – see 4.1.1).



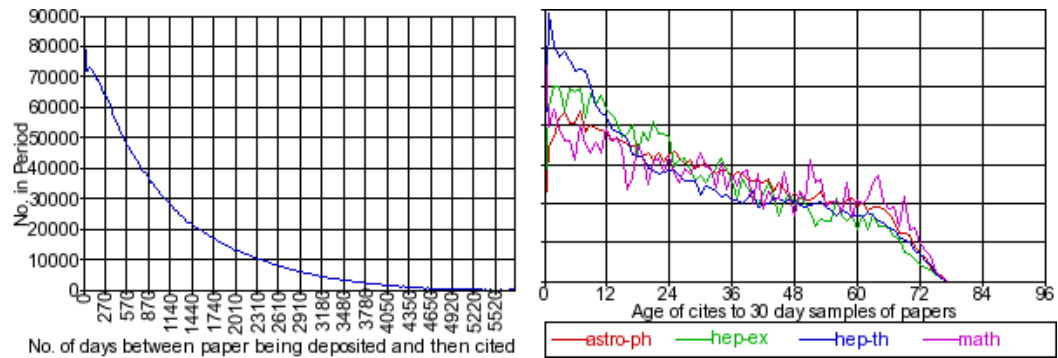
**Figure 7.17:** Histogram of all papers by citation impact (left) and papers rank-ordered by citation impact plotted on logarithmic scales (right)

### 7.6.3 Citation Latency: `citation_latency`, `reference_latency` (per\_area, per\_year)

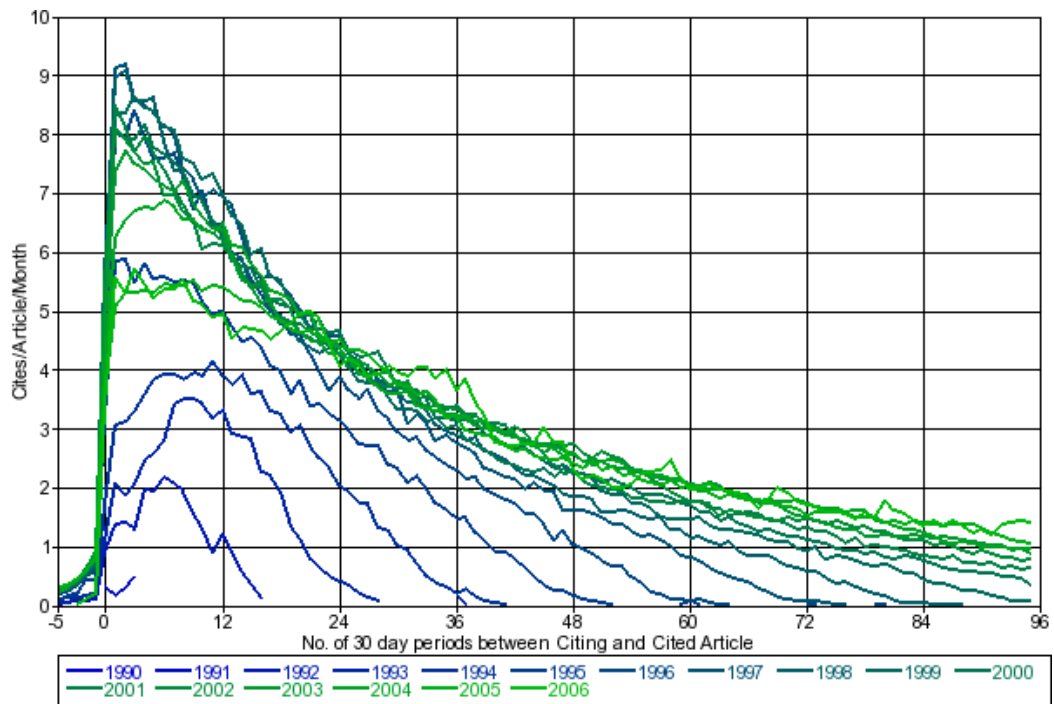
Citation latency is the time-difference between a paper and a citing paper being deposited or published. The statistics script provides four latency analyses: for all papers, by arXiv subject area, by the year of the *citing* paper or by the year of the *cited* paper.

**Figure 7.18** shows the citation latency for all papers (left). The right figure shows the citation latency for all papers for four example arXiv subject areas, normalised by the size of the sample (see subsection 1.4.1 for arXiv abbreviation definitions). The x-axis is the number of days (left) or ‘months’ (right, actually days divided by 30). The y-axis is the number of citations sharing the same citation latency.

The `citation_latency_per_year` is discussed in some depth in section 9.3 (see **Figure 9.9**, page 151). `reference_latency` (**Figure 7.19**) is the same as the per-year breakdown, but by the year of the *cited* paper, rather than *citing* paper.



**Figure 7.18:** Reference latency for all papers (left) and broken down by arXiv subject (right)



**Figure 7.19:** Citation latency per year (based on year of accession of the cited paper)

#### 7.6.4 Age of Cited Papers: `cited_age`

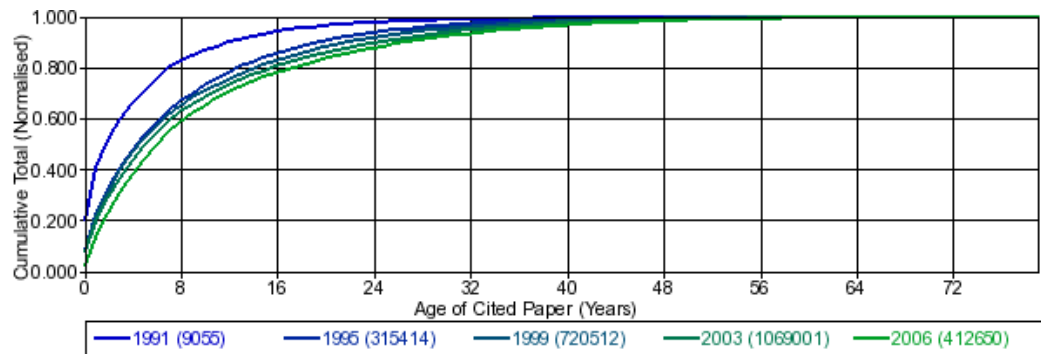
This analysis (*e.g.* Figure 7.21) provides the normalised distribution of the age of cited papers. A sample of five years – spanning the lifetime of the arXiv – are shown. The references are extracted from each years' papers, and the year of the cited publication taken. Because the vast majority of references include at least

the year the cited paper was published this allows the analysis to cover many more citations than are linked by Citebase and to papers published before the arXiv *e.g.* see [Figure 7.20](#). The cited years are then normalised (by dividing by the total number of cited items) and plotted on a cumulative graph.

- |     |  |
|-----|--|
| [1] | H.S. Snyder, "Quantized Space-Time," Phys. Rev. <b>71</b> ( <b>1947</b> ) 38; "The Electromagnetic Field In Quantized Space-Time," Phys. Rev. <b>72</b> ( <b>1947</b> ) 68.                                    |
| [2] | A. Connes, "Noncommutative Geometry," Academic Press ( <b>1994</b> ).  |
| [3] | A. Connes and M. Rieffel, "Yang-Mills For Noncommutative Two-Tori," in Operator Algebras and Mathematical Physics (Iowa City, Iowa, 1985), pp. 237 Contemp. Math. Oper. Alg. Math. Phys. 62, AMS <b>1987</b> . |
| [4] | A. Connes, M. R. Douglas, and A. Schwarz, "Noncommutative Geometry and Matrix Theory: Compactification On Tori," JHEP <b>9802:003</b> ( <b>1998</b> ), <a href="#">hep-th/9711162</a> .                        |
| [5] | T. Banks, W. Fischler, S.H. Shenker and L. Susskind, "M-Theory as a Matrix Model: A Conjecture," Phys. Rev. <b>D55</b> ( <b>1997</b> ) 5112, <a href="#">hep-th/9610043</a> .                                  |

**Figure 7.20:** Plotting the cited age uses all references, because it only requires the year of publication (highlighted in in red)

The shape of the cumulative line gives an indication of the obsolescence of the cited literature. As the shape of the line has flattened during the lifetime of arXiv, this suggests the length of time papers are cited for is increasing. The age of papers cited by arXiv e-prints is discussed in greater depth in [section 9.4](#).

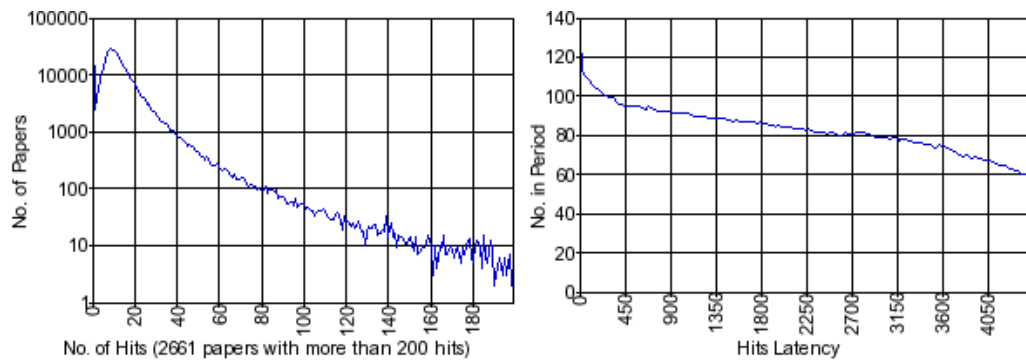


**Figure 7.21:** Year of publication of cited papers, by year of the citing arXiv paper

### 7.6.5 Downloads Analysis: hits\_frequency, hits\_latency, hitsbydomain

The `hits_frequency` and `hits_latency` ([Figure 7.22](#)) show respectively the distribution of papers according to the number of downloads per paper and age of

downloads. NB unlike citations almost all arXiv papers receive some downloads but otherwise share a similar logarithmic distribution. The location of users of the UK arXiv mirror is summarised in [Table 7.1](#) – over 75% of users access the mirror from UK hosts (GB – Great Britain – is UK, US is United States, *etc.*).



**Figure 7.22:** Histogram of arXiv papers deposited in 2000 by the number of downloads to each paper (left) and histogram of hits by age of downloaded paper (right)

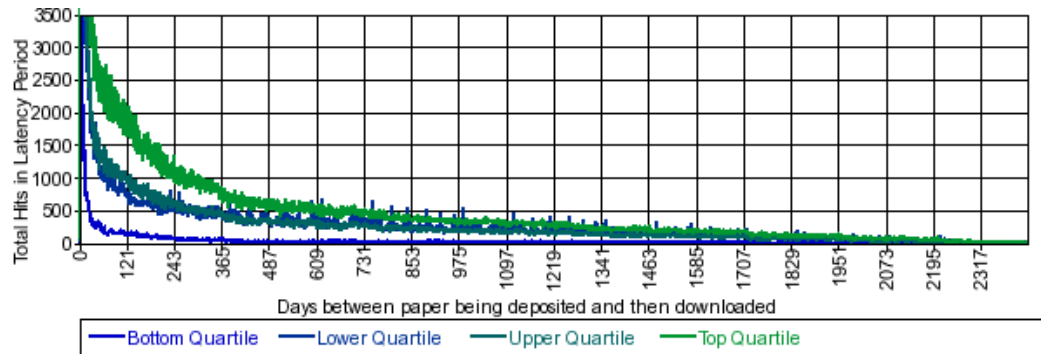
**Table 7.1:** Downloads by country between 2005-07 and 2006-07 (1257768 total).

| Country     | Downloads |
|-------------|-----------|
| UK          | 1022009   |
| USA         | 52863     |
| Bulgaria    | 27533     |
| Japan       | 17705     |
| China       | 13277     |
| Germany     | 11344     |
| Netherlands | 11100     |
| Ireland     | 7623      |
| Poland      | 7364      |
| France      | 5779      |

### 7.6.6 Downloads-Citations Latency Comparison: `hits_latency_normalised`, `hitslatencybyquartile`

The next chapter studies the degree to which downloads and citations co-vary (*i.e.* their correlation). `hits_latency_normalised` shows a normalised comparison of the download and citation latencies (see [Figure 8.3](#), page 127). `hitslatencybyquartile` compares the number of downloads to papers depending

on the papers' citation impact (Figure 7.23); the papers are rank-ordered by citation impact, split into four equal samples (top, upper, lower, bottom quartiles) and plotted as a histogram of number of hits per time latency (in days).



**Figure 7.23:** Histogram of papers by download impact, separated into quartiles by citation impact

### 7.6.7 The Correlation Generator

The Correlation Generator<sup>6</sup> (section 8.8) uses citation and download data from the Citebase Search database to calculate the correlation (and predictive power) of citations and downloads.

## 7.7 Citebase Search Usage Analysis

The awstats<sup>7</sup> tool is used to analysis the usage of the Citebase Search service. Figure 7.24 shows awstat's summary for 2005. Awstats groups together multiple requests into 'sessions', with each unique IP in a session constituting a 'unique visitor'.

Citebase Search got between 100,000 and 180,000 unique visitors per month (excluding November 2005<sup>8</sup>), corresponding to some 1.2-2.6 million hits per month. This accounts for some 1GB of data transferred daily, with another 7GB being harvested by web crawlers (of which 6GB is Google!). Web crawlers

<sup>6</sup>Correlation Generator <http://www.citebase.org/analysis/correlation.php>

<sup>7</sup>AWStats is a generic web log analysis tool, available from <http://awstats.sourceforge.net/>

<sup>8</sup>A fire interrupted service for much of November 2005.



account for roughly 3.5 times the amount of traffic generated by humans (as differentiated by awstats). While Citebase Search makes its usage statistics publicly available, similar statistics aren't available for other OAI based service providers with which to draw a comparison.

| Month    | Unique visitors | Number of visits | Pages    | Hits     | Bandwidth |
|----------|-----------------|------------------|----------|----------|-----------|
| Jan 2005 | 130505          | 200303           | 865852   | 1583212  | 40.98 GB  |
| Feb 2005 | 158026          | 241145           | 806082   | 1669849  | 32.25 GB  |
| Mar 2005 | 168016          | 261000           | 1523341  | 2583309  | 47.56 GB  |
| Apr 2005 | 179622          | 276480           | 841889   | 1906046  | 37.92 GB  |
| May 2005 | 172540          | 271760           | 870056   | 1876358  | 38.36 GB  |
| Jun 2005 | 153237          | 245434           | 952368   | 1843765  | 41.93 GB  |
| Jul 2005 | 112224          | 176595           | 596371   | 1211945  | 26.72 GB  |
| Aug 2005 | 103048          | 171912           | 718690   | 1286383  | 53.87 GB  |
| Sep 2005 | 128779          | 214839           | 1535156  | 2189717  | 77.08 GB  |
| Oct 2005 | 158069          | 263360           | 951218   | 1738399  | 75.68 GB  |
| Nov 2005 | 22979           | 36457            | 166754   | 322682   | 8.12 GB   |
| Dec 2005 | 82940           | 131431           | 745906   | 1238504  | 54.45 GB  |
| Total    | 1569985         | 2490716          | 10573683 | 19450169 | 534.92 GB |

**Figure 7.24:** Citebase Search usage in 2005 (excludes web crawlers and robots).

Given the high level of use of Citebase Search it is perhaps a little disingenuous to describe Citebase Search as an 'experimental' system but, while its scope is limited, it does demonstrate a clear demand for open access citation indices.

Usage statistics aren't published by comparable services.

## 7.8 Conclusion

Citebase Search, a scientometric tool developed at the University of Southampton to explore and demonstrate the potential of an open access corpus, currently harvests open access full-texts from two centralised open access repositories, arXiv and Cogprints<sup>9</sup>, two University of Southampton institutional GNU e-prints repositories (University and the Electronics and Computer Science Department), one publisher-based open access repository, Biomed Central, and several other GNU-eprints repositories (*e.g.* E-LIS<sup>10</sup>).

Citebase Search parses the reference lists from these papers, linking those references that can be found in its database (*i.e.* internal references to the

<sup>9</sup>Cogprints <http://cogprints.org/>

<sup>10</sup>E-LIS <http://eprints.rclis.org/>

repositories it harvests from). The linked references create a ‘citation database’, which allows citation impact – the number of citations to papers or authors – to be counted. Citebase Search can be used to display search results rank-ordered on the basis of the citation counts of either the (1) retrieved papers or the (2) retrieved papers’ authors. Users can search and navigate within this entire full-text corpus via citations to/from each paper, via co-cited papers, via ‘hubs and authorities’ and on the basis of graphs of each paper’s download and citation history.

Usage (download) data is harvested from the UK arXiv mirror and the University of Southampton repositories. Whether usage data can be used to predict later citation impact is analysed in the next chapter.

# Chapter 8

## Using Web Statistics for Usage Analysis

### 8.1 Introduction

The use of citation counts to assess the impact of research papers is well established. Because citations come from other papers, there is a lead-time between a paper being published, read and then cited by other authors (and for this to happen a sufficient number of times to be statistically useful), hence the citation impact of a paper can only be measured several years after it has been published. To date citations have been the only evidence of a paper's use but as research papers are increasingly accessed through the web, the number of times a paper is downloaded can be instantly recorded and counted<sup>1</sup>. One might expect the number of times a paper is read to be related both to the number of times it is cited and to how old the paper is. This chapter analyses how well short-term web usage impact predicts medium-term citation impact. Citebase Search is used to test this (restricted to the arXiv collection of papers).

---

<sup>1</sup>The number of downloads is assumed to approximate the number of reads, once automated web crawlers are excluded.

### 8.1.1 Why predict citation counts?

Peer-reviewed journal paper (or refereed conference paper) publication is the primary mode of communication and record for scientific research. Researchers – as authors – write papers that report experimental results, theories, reviews, and so on. To relate their findings to previous findings, authors cite other papers. Authors cite a paper if they (a) know of the paper, (b) believe it to be relevant to their own paper and (c) believe it to be important enough to refer to explicitly (*i.e.* there is both a relevance and an importance judgment inherent in choosing what to cite). It is probably safe to assume that the majority of citations will be positive, but even negative citations (where an author cites a paper only to say it is wrong or to disagree with it) will refer to papers that the author judges relevant and important enough to warrant rebuttal [Borgman and Furner \(2002\)](#) provide a review of many studies that debate the motivations for and influences on citing. Citations can therefore be used as one measure of the importance and influence of papers, as well as indirectly the importance of the journals they are published in and the authors that wrote them. The total number of times a paper is cited is called its *citation impact*.

The time that it takes – from the moment a paper is accepted for publication (after peer review) until it is (1) published, (2) read by other authors, (3) cited by other authors in their own papers, and then (4) those citing papers are themselves peer-reviewed, revised and published – may range anywhere from 3 months to 1-2 years or even longer (depending on the field, the publication lag, the accessibility of the journal, and the field’s turnaround time for reading and citation). In physics, the “cited half-life” of a paper (the point at which it has received half of all the citations it will ever receive) is around 5 years: [ISI \(2004\)](#) shows most physics-based journals having a cited half-life between 3 and 10 years. Although papers may continue to be cited for as long as their contents are relevant (in natural science fields this could be forever), citation counts using the ISI Journal Impact Factor ([Garfield, 1994](#)) use only 2 years of publication data in a trade-off between (i) a paper being recent enough to be useful for assessment and (ii) allowing sufficient time for sufficient citations to accrue to be statistically meaningful.

### 8.1.2 Web accesses as an early-day predictor

Is it possible to identify the importance of a paper earlier in the read-cite cycle, at the point when authors are accessing the literature? Now that researchers access and read papers through the web, every download of a paper can be logged. The number of downloads of a paper is an indicator of its usage impact, which can be measured much earlier in the reading-citing cycle.

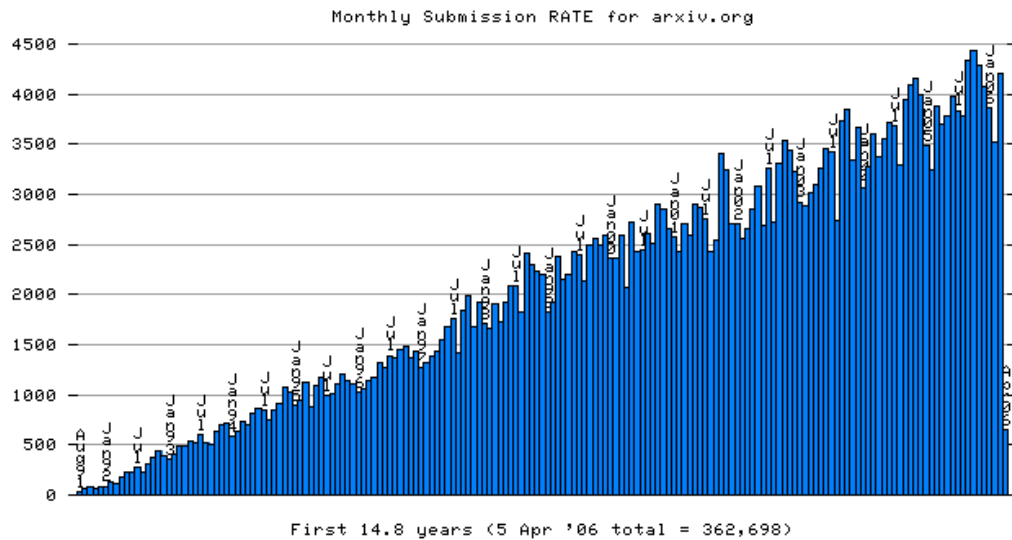
This chapter uses download data from the UK mirror of arXiv and citation data from Citebase Search to test whether early usage impact can predict later citation impact. Since I started work on correlating download and citation data several other similar studies have been performed, including [Bollen et al. \(2005\)](#); [Mayr \(2006\)](#); [Moed \(2005\)](#). The closest work to my own is [Perneger \(2004\)](#), who performed a study based on the British Medical Journal (Perneger used web accesses to the journal's web site and citation data from the ISI Web of Science).

## 8.2 Chapter Structure

The following section describes the arXiv e-print archive and the data used from its UK mirror for this study. I describe how the citation data is constructed in Citebase Search (described in the previous chapter), an autonomous citation index similar to CiteSeer (see [subsection 3.3.2](#)). I introduce the Usage/Citation Impact Correlator, a tool for measuring the correlation between paper download and citation impact. Using the Correlator I have found evidence of a significant and moderately large correlation between downloads and citations. I accordingly conclude that downloads can be used as early-days predictors of citation impact.

## 8.3 arXiv

arXiv is an online database of self-archived research papers covering physics, mathematics, and computer science ([Ginsparg, 2003](#)). Authors deposit their papers as preprints (before peer review) and postprints (after peer review – both referred to here as “e-prints”) in source format (often LaTeX), which can be converted by the arXiv service into postscript and PDF. In addition to depositing

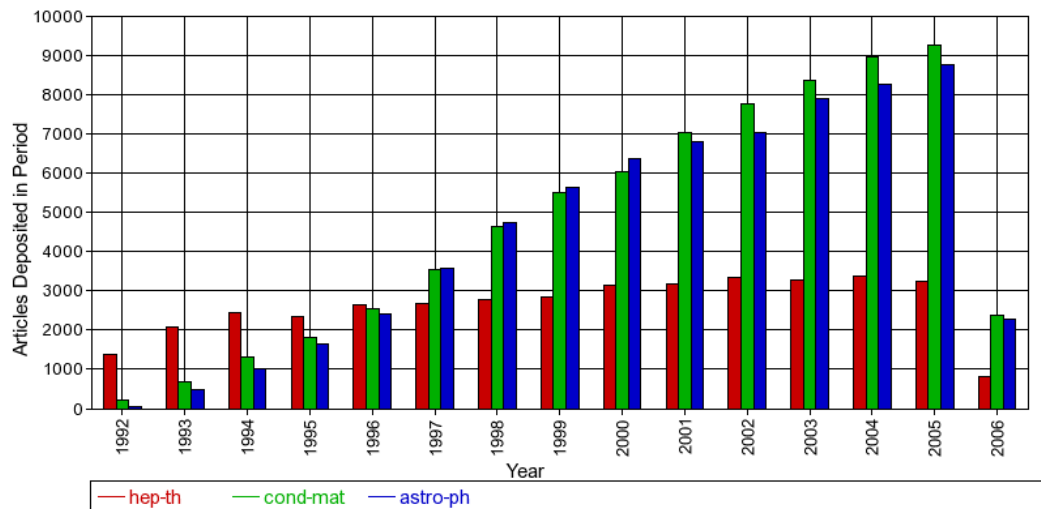


**Figure 8.1:** Monthly submission rate for arXiv (source arXiv)

the full-text of the paper, authors provide metadata. The metadata include the paper title, author list, abstract, and optionally a journal reference (where the paper has been or will be published). Papers are deposited into “sub-arXivs”, subject categories for which users can receive periodical email alerts listing the latest additions.

The annual number of papers deposited in arXiv has been growing at an linear rate since 1991 (see [Figure 8.1](#)). Hence, assuming that the total number of papers written each year is relatively stable, arXiv’s coverage is also increasing linearly (the proportion of all potential papers that could be deposited in arXiv, but haven’t been). This may not be true of all sub-areas of arXiv however. [Figure 8.2](#) shows the number of new records in several sub-fields. The High Energy Physics (HEP) sub-area is growing least (because most of the material within that arXiv subject is already being self-archived), whereas Condensed Matter and Astrophysics are still growing considerably. [Kurtz et al. \(2005\)](#) found 74% of papers published by the Astrophysical Journal in 2003 also had a version deposited in arXiv, suggesting that even within the Astrophysics sub-area at least one journal is almost fully ‘arXived’. As HEP is an older arXiv area than Astro it is likely those journals whose papers fall within arXiv’s HEP field will have higher percentages self-archived in arXiv.

In addition to being aided by the wide coverage of the HEP sub-arXiv, Citebase’s ability to link references in the HEP field is strengthened by the addition of the



**Figure 8.2:** Recent growth in arXiv is due to Cond-Mat and Astro-Ph

journal reference to arXiv's records by SLAC/SPIRES. SLAC/SPIRES indexes the table of contents from HEP journals by hand, and links the published version to the self-archived e-print version in arXiv (if available), adding the journal reference to the arXiv record. Where an author cites a published paper without providing the arXiv identifier, Citebase can use the data provided indirectly by SLAC/SPIRES to link that citation, thereby counting it in the citation impact. arXiv doesn't provide metrics on the number of author versus SLAC/SPIRES supplied journal references.

With 360,000 papers self-archived over fifteen years, arXiv is the largest self-archived centralised e-print archive. There exist bigger archives, such as Citeseer whose contents are computationally harvested from distributed sites. The Astrophysics Data Service (ADS) by scanning back catalogues and in collaboration with publishers provides comprehensive free-access to the Astrophysics literature. arXiv is an essential resource for research physicists, receiving 10,000 downloads per hour from the main mirror site alone, supplementing this main point of access are a dozen mirror sites, of which the UK mirror is located at the University of Southampton.

## 8.4 Harvesting from arXiv

ArXiv provides access to its metadata records through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in Dublin Core format. As the full-texts are available without restriction, these are harvested by a web robot (which knows how to retrieve the source and PDF versions from arXiv's web interface). Both metadata and full-text are stored in a local cache at Southampton.

Web logs in Apache “combined” format are sent from the UK arXiv mirror server via email and stored locally. (Requests for web logs from the other arXiv mirror sites, including the main site in the US, have not been granted.) The web logs are filtered to remove common search engine robots, although most web crawlers are already blocked by arXiv. Requests for full-texts are then extracted *e.g.* URLs that contain “/pdf/” for PDF requests. On any given day only one full-text download of a paper from one host is counted (so one user who repeatedly downloads the same paper will only be counted once per day). This removes problems with repeated requests for the same paper, but results in undercounts when more than one user requests a paper from a single host or from behind a shared network proxy. This study cannot count multiple readings from shared printed copies, nor readings from copies in other web distribution channels such as publishers' web sites.

Each full-text request is translated to an arXiv identifier and stored, along with the date and the country of the requesting host (*e.g.* “UK”). This corresponds to some 4.7 million unique requests from the period August 1999 (when the UK arXiv mirror was set up) to October 2004. Because only one mirror's logs are available, this biases the requests towards UK hosts (see [Table 7.1](#), page 112) and possibly as a result towards UK-authored papers. This bias cannot be tested or corrected unless the logs are made available from other mirrors and augmented with data from other e-print archives – as I hope these results will encourage them to be!



## 8.5 Citebase

Citebase is an autonomous citation index. Metadata records harvested from arXiv (and other OAI-PMH archives) are indexed by Citebase. The full-texts from arXiv are parsed by Citebase to extract their reference lists. These reference lists are parsed, and the cited papers are looked up in Citebase. Where the cited paper is also deposited in arXiv, a citation ‘link’ is created from the citing paper to the cited paper (stored in the database as pairs of paper identifiers). These citation links create a citation database that allows users to follow links to cited papers (“outlinks”) and to see what papers have cited the paper they are currently viewing (“inlinks”).

The total number of citation inlinks to a paper provides a citation impact score for that paper. Within Citebase the citation impact – as well as other performance metrics – can be used to rank papers when performing a search.

The citation impact score found by Citebase is therefore dependent upon several systematic factors: whether the cited paper has been self-archived, the quality of the bibliographic information for the cited paper (*e.g.* the presence of a journal reference), the extent to which Citebase was able to parse the references from citing papers, and how well the bibliographic data parsed from a reference matches the bibliographic data of the cited paper. Citebase’s citation linking is based either upon an arXiv identifier (if provided by the citing author), or by bibliographic data. Linking by identifier can lead to false positives, where an author has something in their reference that looks like an identifier but isn’t, or where an author has made a mistake (in either case the reference link goes to the wrong paper). Linking by bibliographic data is more robust, as it requires four distinct bibliographic components to match (author or journal title, volume, page and year), but this will obviously be subject to some false positives (*e.g.* where two references are erroneously counted as one) and uncounted citations from missed links.

## 8.6 Accuracy of Citation Links within Citebase

The citation impact of papers within this study is dependent upon the number of references found by Citebase to those papers. An absolute limit on the number of citations is the coverage of the body of literature analysed, *i.e.* arXiv's holdings. The question of coverage is a general problem for any study of citations, as anything from the references from a single journal, to links from web pages may be counted. Of course, there could be significant differences in the purpose of making a citation between subjects, journals, or the web in general. Within the context of this study, we are comparing downloads and citations from the same source, so the biased coverage of citations (only from arXiv papers to arXiv papers) will also be a bias in downloads (arXiv downloads only).

No references have been successfully extracted from 7000 (2%) arXiv papers. This may arise because of the document format and reference style. It is an ongoing optimization process to keep trying to decrease the number of papers for which no references can be extracted, and to increase the accuracy of that extraction.

Papers that are published in a journal have two manifestations that may be cited by authors – the arXiv e-print and the 'official' version available from the publisher's website. The arXiv version may be cited using the arXiv identifier, either in the absence of or in addition to the bibliographic reference to the journal. The publisher's version may be cited only by the bibliographic reference (particularly where an arXiv version may be deposited at a later date as a post-print).

To link a bibliographic journal reference the bibliographic data needs to be available for the cited paper: the journal title, volume, pagination *etc.*

Unfortunately many authors do not provide the journal reference for the published paper that they have also deposited as an e-print in arXiv. This means that any citations that do not include the arXiv e-print identifier are missed, potentially reducing the number of citations identified by Citebase to a paper.

To check the accuracy of Citebase's reference linking a sample of 500 randomly selected papers was chosen, spread across all years of the arXiv (1991-). 90 papers from 2003-2005 were checked by hand to ascertain the number of references that

could be linked (because the cited paper is in Citebase) but wasn't.

Reference lists on Citebase's abstract pages provide several web links – depending on the available data – to help find cited papers: a query to Google Scholar using authors' names, journal title and publication year, for some publishers a link to the journal paper and for this study an additional link was created to query Citebase using the authors' names and publication year. It was assumed that papers older than 1992 would not be in arXiv (arXiv started in 1991) nor would arXiv contain any cited books.

Over time Citebase's capabilities have been extended, so the initial search in this study for the cited item was made using the latest iteration of Citebase's OpenURL link resolver (that makes use of additional rules to tidy up references, which are run only occasionally against the entire database). If the direct Citebase lookup failed, queries to Google Scholar or the publisher were made to ascertain a fuller citation for the cited item (in particular to get the title of the cited paper, which is rarely included by physicists in references). If a title was forthcoming from either of those sources it was used to query Citebase, otherwise a query based upon only authors and year was made with the cited item. Where an unlinked reference couldn't be found in Google Scholar or by publisher (*i.e.* an unknown journal) it was assumed that it was not in Citebase.

During this study it became clear that a common cause of link failure was inconsistent formatting of journal and volume, *e.g.* "Phys.Rev.D65" and "Phys.Rev.D 65", causing the 'D' to be picked up as part of the volume or journal title respectively. Citebase has since had additional rules added to re-format these cases consistently, although to keep this study consistent these additional citation links were not included.

**Table 8.1** shows a summary of the 500 papers studied. Around 40% of references were successfully linked to the cited paper by Citebase, but a total of 64 papers had no references linked. The authors of the sample paper analysed in **Table 8.2** had already provided arXiv e-print identifiers for all references that I could locate in arXiv (*i.e.* Citebase couldn't provide any more citation links than the authors have already provided). Two 'references' were actually a single reference that had been split into two by Citebase around a semi-colon (treated by Citebase as a separator between references), so the actual number of authored references is 63. 'Misc. material' are what looks like technical reports, while 'non-published' is *e.g.*

**Table 8.1:** Summary of sample set used to test Citebase's reference parsing and citation linking accuracy

|   |     |
|---|-----|
| Total Papers                                    | 500 |
| Papers without any references                   | 8   |
| Papers with no linked references (inc. no refs) | 64  |
| Average number of references/paper              | 33  |
| Average number of references linked/paper       | 14  |

**Table 8.2:** Summary of reference links from an example paper.

|   |                              |
|---|------------------------------|
| Paper Identifier                        | oai:arXiv.org:hep-ph/0502036 |
| References                              | 64                           |
| References w/arXiv identifier           | 48                           |
| Reference links missed to arXiv papers  | 0                            |
| References unlinked w/journal reference | 8                            |
| References to misc. material            | 5                            |
| References to non-published items       | 3                            |

‘private communication’ (hence unlikely to ever be linkable).

The most recent 90 papers ([Table 8.3](#)) from the random selection of 500 papers were chosen to look at in detail. A total of 1293 references were found that weren't linked to the cited item, although most of those 1293 references were to items older than arXiv or not to journal papers. On average 1.8 references were found in each paper that were in arXiv, but not linked by Citebase *i.e.* 5% of all references aren't linked by Citebase, even though both the citing and cited paper are in arXiv.

**Table 8.3:** Most recent 90 papers

|  |       |
|--|-------|
| 2003-2005 Papers                                       | 90    |
| Average number of references/paper                     | 35.26 |
| Average number of references linked/paper              | 21.46 |
| Average number of references failed to be linked/paper | 1.80  |

## 8.7 Correlation between Citations and Downloads

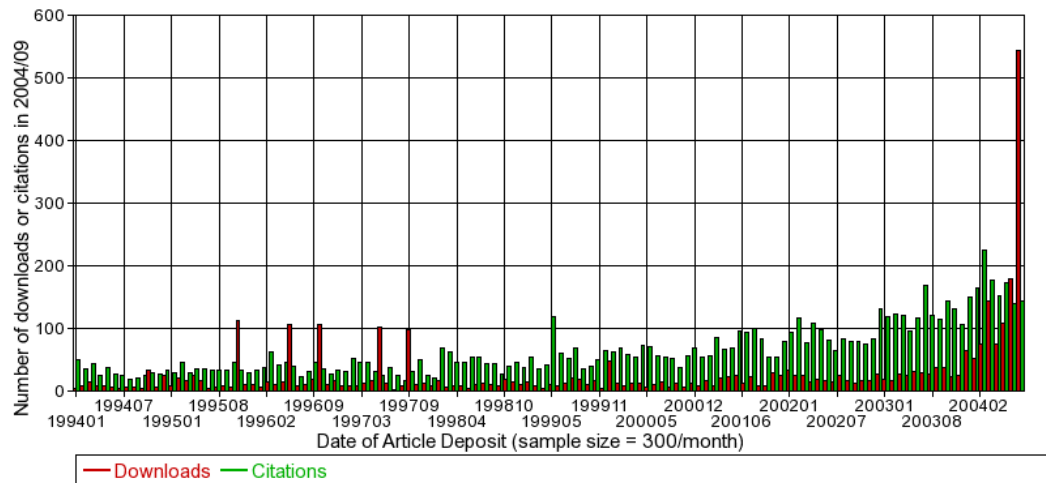
Correlation is a statistical measure of the degree to which two variables co-vary. Two positively correlated variables  $x$  and  $y$  will tend to have high values of  $x$  paired with high values of  $y$ , and low values of  $x$  with low values of  $y$ . A negative correlation is where high values of  $x$  are paired with low values of  $y$ . Correlation is a normalised, scale-independent measure based on standard deviations above and below each variable's mean – the raw values of  $x$  and  $y$  can be in any number range.

A correlation between  $x$  and  $y$  may occur because  $x$  influences  $y$ ,  $y$  influences  $x$ ,  $x$  and  $y$  influence each other, or an external variable influences both  $x$  and  $y$ .

Intuitively, one would expect citations and downloads to exert a bi-directional influence, cyclical in time: An author reads a paper A, finds it useful and cites it in a new paper B (download causes citation). Another author reads B, follows the citation, reads A (citation causes download) and then perhaps goes on to cite it in another paper, C (download causes citation), *etc.* The correlation will be less than 1.0, not only because we don't cite everything we read, nor read everything that a paper we read cites, but because both downloads and citations are subject to other influences outside this read-cite-read cycle (*e.g.* from alternative discovery tools, or when authors copy citations from papers they read without reading the cited works – perhaps when they have read the paper before).

Reader-only users contribute to the cite-read effect but not the read-cite effect, adding further noise to the read-cite effect.

Monitoring the correlation between citations and downloads is also informative because although papers can be downloaded and cited for as long as they are available, the peak rates for downloads and citations tend to occur at different times. [Figure 8.3](#) shows a histogram of papers downloaded (red bars) or cited (green bars) in September 2004 by age of the downloaded/cited paper. Papers in arXiv that are over a year old show an almost flat rate of downloads whereas their citation rate shows a more linear rate of decay over the period of available data. The most frequently downloaded papers are those deposited in the previous year (2003-) with a steep fall-off during that year; for papers from earlier years downloads are (with a few exceptions) few and equal from year to year. For



**Figure 8.3:** Age of papers downloaded or cited in 2004/09 (normalised).

citations the fall-off looks more gradual and linear, taking about six years or more to settle into a flat, constant rate. If higher impact papers account for that higher rate of downloads in the first year, then the initial year of download data could be used to predict citation impact data over the longer term.

Figure 8.3 was generated by taking a random sample of 300 papers from those deposited each month from January 1994 until September 2004. Each sample month is plotted as a bar (downloads in red and citations in green). By taking a fixed size sample of papers this analysis normalises for the changing size of arXiv.

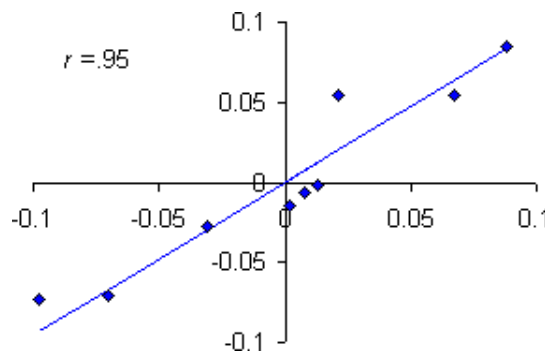
If there is a correlation between citations and downloads, a higher rate of downloads in the first year of a paper could predict a higher number of eventual citations later. To test this I built a “Correlation Generator” to analyse the relationship between the citation and download counts for research papers in arXiv and to test whether a higher rate of downloads leads to a higher rate of citations.

## 8.8 Correlation Generator

The correlation generator calculates the *Pearson’s Product Moment Correlation* (Pearson’s  $r$ ) and *Spearman’s Rank Correlation* between citations and downloads

(or citations-citations, downloads-downloads<sup>2</sup>). A scatter plot is also generated that provides a visual representation of the distribution of citations and downloads.

Pearson's  $r$  is the degree of co-variance between two variables. If  $r$  is  $-1$  or  $1$  then  $x$  and  $y$  co-vary exactly or if  $r$  is  $0$  then there is no relationship between  $x$  and  $y$ . Whether  $r$  is positive or negative indicates whether  $y$  increases or decreases as  $x$  increases.  $x$  and  $y$  must be linearly related, that is when the data points are plotted on an  $xy$  plot the points should vary around a line of the form  $y = mx + c$  e.g. see Figure 8.4. Because the distributions of citations and downloads are skewed the natural logarithm is used to make the data points closer approximate a normal distribution around the line of best fit (compare the plots in Figure 8.5, without logarithm, and Figure 8.6, with logarithm).



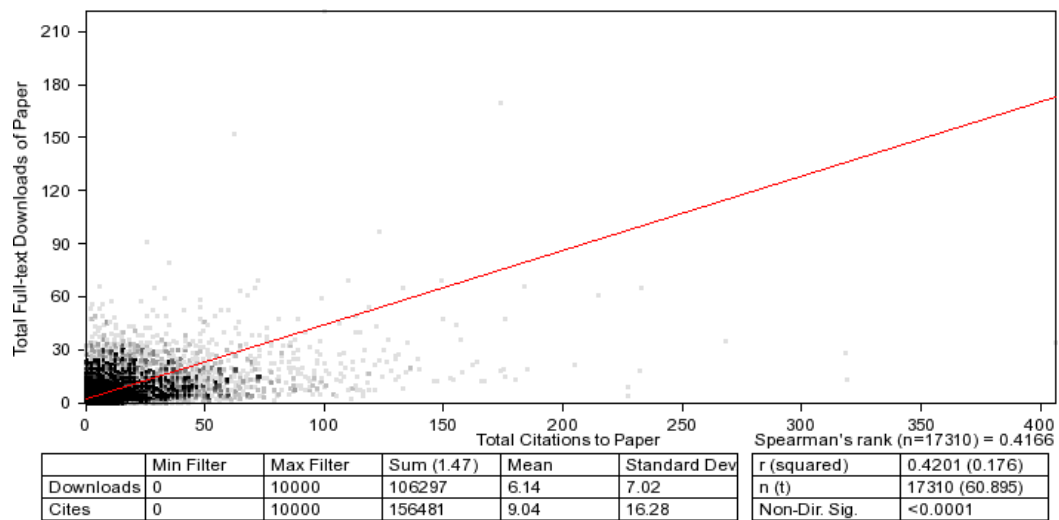
**Figure 8.4:** Example distribution with Pearson's  $r$  line plotted

Spearman's Rank correlation first rank-orders  $x$  and  $y$  and then finds Pearson's  $r$  of the two rank-orders. Spearman's Rank can be used on any distribution of ordinal data (*i.e.* values that can be ranked). In this chapter I have used Pearson's  $r$  in preference to Spearman's Rank.

Spearman's Rank correlation first rank-orders  $x$  and  $y$  and then finds Pearson's  $r$  of the two rank-orders. Spearman's Rank can be used on any distribution of ordinal data (*i.e.* values that can be ranked). In this chapter I have used Pearson's  $r$  in preference to Spearman's Rank.

The correlation generator provides a number of filters that can be used to restrict the data going into the correlation. As the correlation is calculated from pairs of

<sup>2</sup>Citation-citation correlation is used to correlate early-day citations with later citation impact.



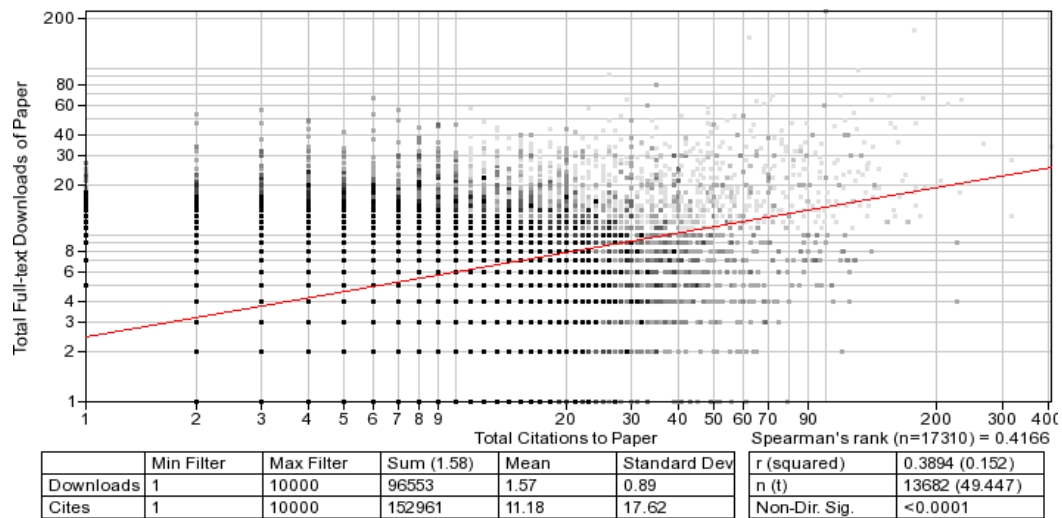
**Figure 8.5:** Correlation scatter graph without logarithmic translation

citation and download counts corresponding to one paper, the filters determine which papers to include and which downloads and citations to those papers to count. The values used can be specified symbolically, by entering values into the form, or graphically, by clicking on the mini-graphs to define upper and lower limits (Figure 8.7).

Figure 8.8 shows the options available in the correlation generator. The user can choose what data and data-ranges are used to generate the correlation between downloads (“Hits”) and citations (“Impact”). This provides filters to delimit which papers (Field, Min/Max Downloads, Min/Max Impact, Date), which downloads and citations (Min/Max downloads latency, Min/Max citation latency), and which citation quartile to include. The citation quartile includes in the result only the bottom, lower, upper, or top 25% of papers after rank-ordering by citation impact. Clicking ‘Generate’ calls a web script that extracts the data sets from Citebase, generates the correlation, and displays the result as a graph.

The graphs shown in Figure 8.7 show the distribution of the variables that go into the correlation (note that most use a logarithmic scale). Citation frequency shows the distribution of papers in terms of the number of times they were cited. Citation latency is the time between a paper being deposited and later cited (total citations per day latency). Web download frequency is the distribution of papers in terms of the number of times they were downloaded. Web download latency is the time between a paper being deposited and later downloaded. The user can click on the graphs to set minimum and maximum values, which are





**Figure 8.6:** Correlation scatter graph with logarithmic translation

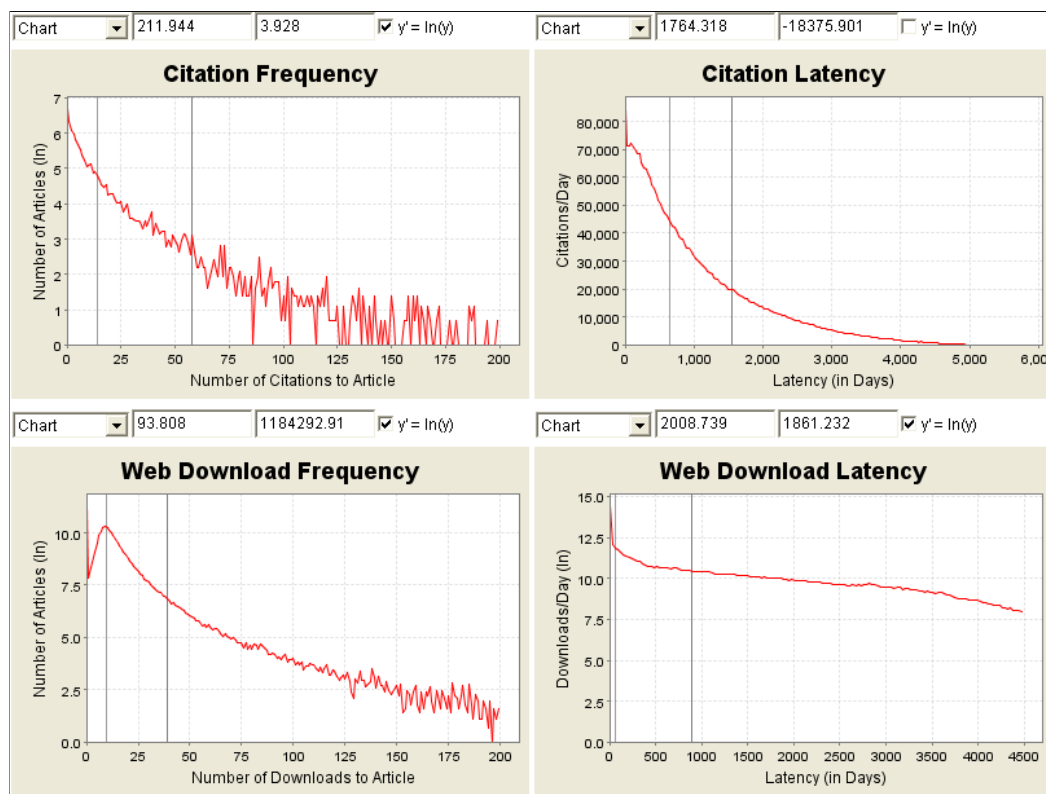
filled into the query form.

Filtering is particularly useful for restricting the analysis to papers for which sufficient data are available: for example, whereas there exist papers that have been deposited since 1991, the download data are only available from 1999. Hence, although the download data cover all of the papers deposited up to that date, the predictive power of downloads can only be tested for papers deposited since 1999.

Each citation and each download has a latency value: (1) the time between when the cited and citing paper were deposited, and (2) the time between when a paper is deposited and when it is downloaded. The user might chose to include only downloads that occurred a week after a paper was deposited (to exclude the initial ‘alerting’ rush of users). By specifying an upper limit to the latency allows meaningful comparisons between papers, because the citations or downloads are restricted to only those downloads to each individual paper upto the given latency period.

## 8.9 Correlation Generator Implementation

The generator is based on Citebase’s MySQL database and a combination of Perl server-side scripts and Java client-side web apps. The tables relevant to



**Figure 8.7:** Preview graphs show the distribution of key variables and allow visual parameter selection

calculating the correlation are the citation links table, download “hits” table, and record timestamps (the earliest timestamp is taken as the date of accession – the date used by the generator to determine latency and to filter by date). The downloads and citation links are pre-processed into latency tables that consist of the paper’s identifier and the number of days since accession, *e.g.* a paper 58432 deposited on the 14th May and then downloaded on the 21st May is stored in the downloads latency table as “58432-7”, similarly 58432 cited by a paper 69710 deposited on 26th May is stored in the citations latency table as “58432-12”.

At the time of writing the download latency table contained 4 million records and the citation latency table 2.3 million records, which have to be processed in real-time to support an interactive tool. Citebase is updated daily with new papers from the source repositories which, in addition to refinements to Citebase’s processes, means the data set used by the correlation generator changes over time.

To provide the user with a graphical representation of the source data the database tables are rendered by Java graphs (Figure 8.7) that retrieve the data as

|                                  |                 |
|----------------------------------|-----------------|
| Database                         | <b>citebase</b> |
| Field                            | All ▾           |
| \$VAR1                           | Downloads ▾     |
| Normalised (Logarithm)           | Yes ▾           |
| Minimum \$VAR1                   | 1               |
| Maximum \$VAR1                   | 10000           |
| Minimum Impact                   | 1               |
| Maximum Impact                   | 10000           |
| Papers Dated From                | 19000000        |
| Papers Dated Until               | 20101231        |
| Downloads Latency Min. (in days) |                 |
| Downloads Latency Max. (in days) |                 |
| Cites Latency Min. (in days)     |                 |
| Cites Latency Max. (in days)     |                 |
| Quartile (by Citations)          | All ▾           |
| Output                           | Graph ▾         |

**Figure 8.8:** Available parameters for customising the correlation calculation

a plain-text list of values from a supporting server-side script. The graphs are tied by client-side Javascript into the main submission form (Figure 8.8) allowing the user to ‘click’ on the graphs, with the appropriate values being filled into the form.

When the user submits the web form the citations and downloads are first filtered to only those papers with identifiers in the given arXiv sub-field, papers whose accession is within the given date range, and only those downloads and citations that occurred within the given latency range. When calculating Pearson’s  $r$  papers with no citations and/or no downloads are discarded (because the logarithm of zero is negative infinity<sup>3</sup>). The citations and downloads are sub-totalled for each paper, from which the logarithm is taken. The correlation is calculated from the citations and downloads values. The server-side script either returns a scatter graph, a summary table, and the correlation or – if the user changes the output to “table” – a list of paired values allowing the matching papers’ citation/download count pairs to be imported into a separate statistical package.

<sup>3</sup>Removing papers with zero citations also removes papers that have no citations due to systematic errors in Citebase.

## 8.10 Generating Correlations

The correlation generator builds a scatter plot, as well as calculating the basic distribution (mean/standard deviation) of the citation and download counts, and the correlation between the two. The scatter plot consists of density dots – the darker the colour the more pairs exist at the same point. This helps to emphasize where the bulk of the pairs lie.

The basic statistical information on the number of pairs used ( $n$ ), and the distributions of the two variables (sum, mean, and standard deviation) is shown. Both citations and downloads have large deviations from the mean, as they are very skewed distributions (*i.e.* a small number of very-high-impact papers account for most downloads and citations, while the majority of papers receive few or no downloads or citations) – as shown in [Figure 8.7](#).

The correlation generator calculates the value for Pearson's  $r$  – the degree to which data points deviate from the line of best fit. Pearson's gives values from -1 to 1, where -1 is a perfect negative correlation, 0 no correlation and 1 a perfect positive correlation. Pearson's in effect provides gradient of the “line of best fit,” which goes through the mean (the correlator draws this line in red). Pearson's is intended to be used for data that is normally distributed, therefore to normalise the distributions for use with Pearson's  $r$  the natural logarithm of downloads and citations is taken; hence the correlator uses a logarithmic axis. The algorithms used by the correlator were checked by entering the same source data into Microsoft Excel (Pearson's only) and SPSS to compare the results for Pearson's  $r$  and Spearman's Rank.

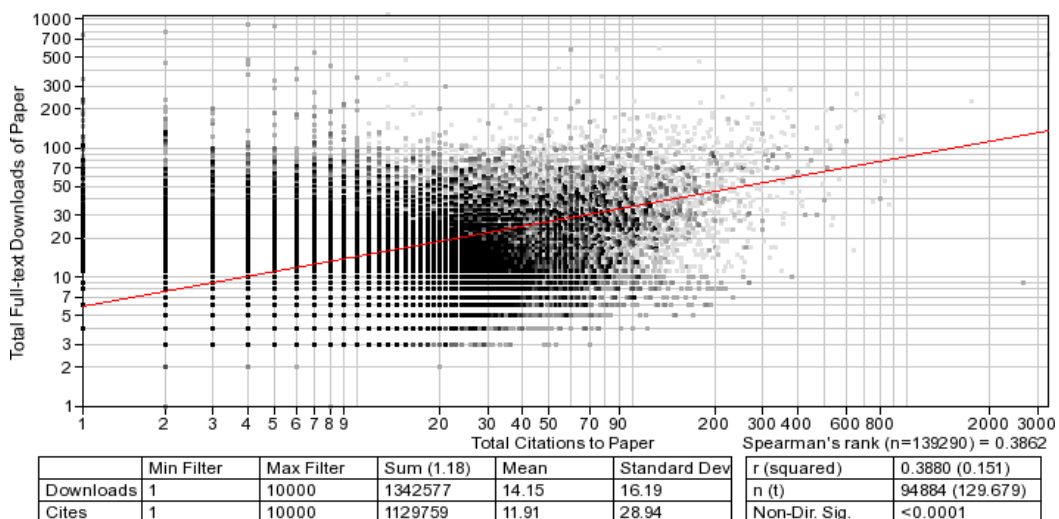
When generating the correlation any downloads within the first 7 days of the paper appearing were excluded, as these downloads reflect users scanning all new papers (*e.g.* in response to email alerts), hence those downloads are unlikely to discern between high impact and low impact papers and would dampen any predictive effect. The first 7 days of downloads accounted for on average .5 downloads per paper – roughly a fifth of all downloads.

While many papers may be downloaded, but not cited, papers that are highly cited are always downloaded. This can be seen in the scatter graphs in the following section; as high-download high-impact papers fall closer to the line of best fit, while low-impact papers appear to form a separate distribution above the

line of best fit. Some papers that have high download counts but low citation counts may be the result of Citebase failing to find the citation links *e.g.* where the author hasn't supplied the journal reference, hence any citations to those papers using only a journal reference cannot be linked and counted.

## 8.11 Sample Correlations

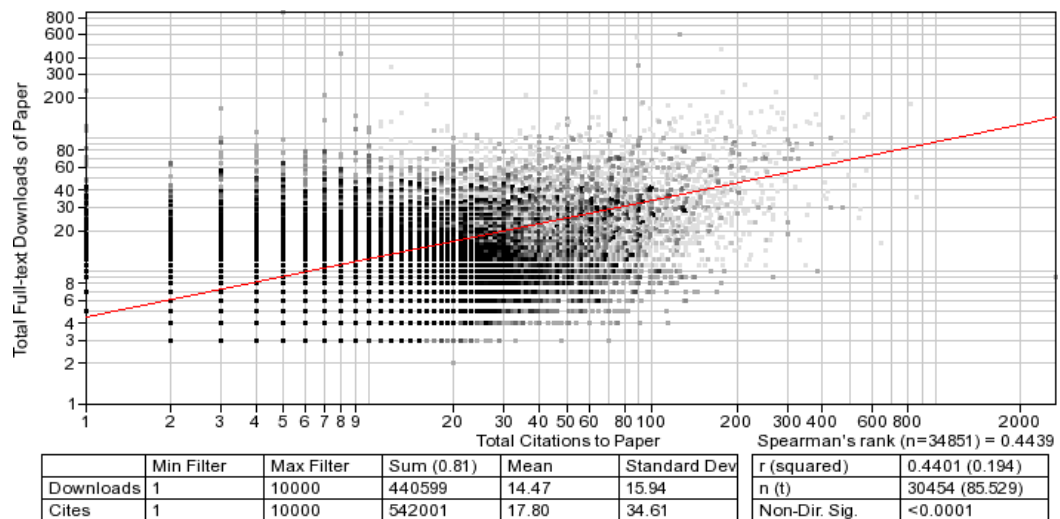
**Figure 8.9** is the download/citation correlation scatter-plot generated for all papers deposited between 2000-2003. Each dot corresponds to a paper. The number of papers with the same values is indicated by shadings of grey (black being the highest, with 4 or more papers having the same number of downloads and citations). The download and citation counts are the cumulative amounts up to April 2006. The correlation for these 94,884 papers is  $r = 0.39$ . From the distribution in the scatter graph it can be seen that the distribution is very noisy, but that papers with high citation impact also receive high numbers of downloads. The ratio of downloads to citations is 1.18:1 (only download statistics for the UK mirror are available), which corresponds to a mean of 11.91 citations for each paper, and 14.15 downloads. Non-Dir. Sig. is the (2-tailed) probability of such a correlation by chance ( $p < .0001$ ).



**Figure 8.9:** Scatter-plot for all papers deposited in arXiv between 2000-2003 (excluding first seven days of downloads)

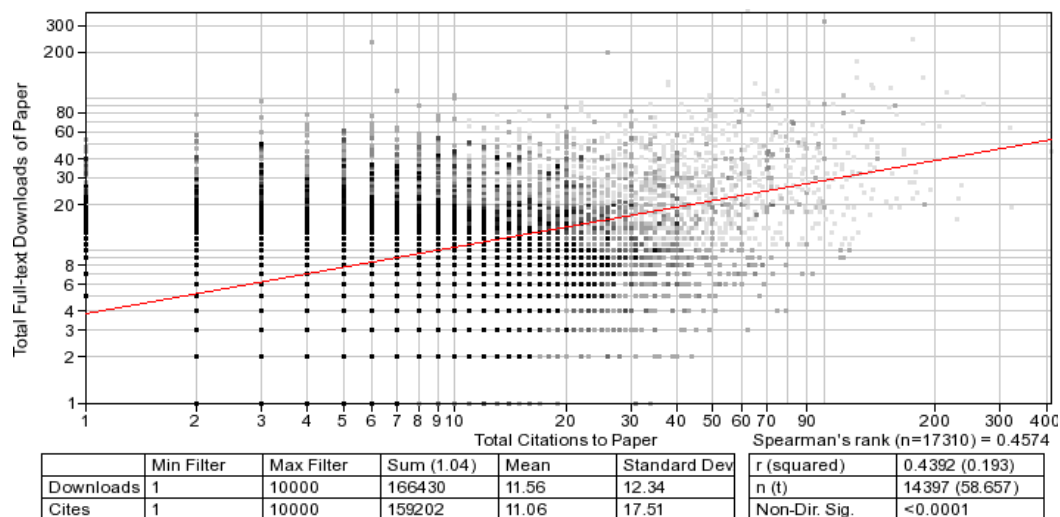
**Figure 8.10** is the correlation for papers deposited in 2000-2003, from the High Energy Physics sub-arXivs (HEP). HEP papers have a mean citation impact of

17.80, and download impact of 14.47. The correlation is higher at  $r = 0.44$  (most likely due to more accurate citation counts for the HEP sub-field).



**Figure 8.10:** Scatter-plot for *hep* papers deposited in 2000-2003

Figure 8.11 is the correlation for papers deposited 2000-2002 (a 2 year period) in HEP. The correlation was also restricted to downloads and citations up to 2 years after the paper was deposited, which while resulting in a slightly lower correlation still rounds to  $r = .44$ .



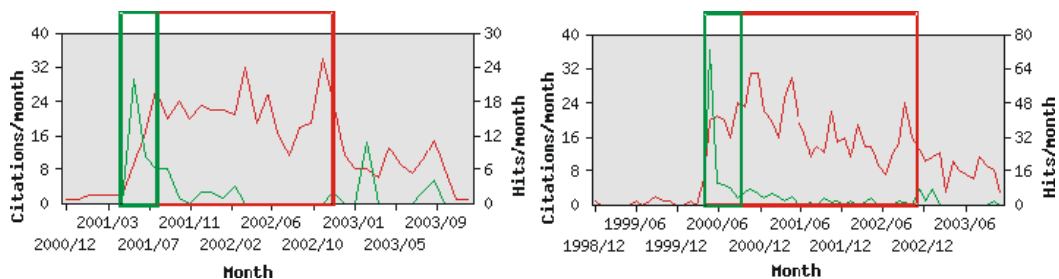
**Figure 8.11:** Scatter-plot for HEP papers deposited in 2000-2001, counting only citations and downloads up to two years after deposit

## 8.12 Predicting citation impact from downloads

The papers used for testing how well downloads can predict citation impact are from the High Energy Physics sub-arXivs and deposited between 2000 and 2002 (a 2 year period). The first 7 days of downloads are excluded, to minimise the effect of the first rush in which all new paper-titles are down-loaded about equally much after appearing in alerting lists.

To test how well downloads predict citation impact, different latency filters are used for the download impact *e.g.* “How well does the download impact after 2 months predict the citation latency after 2 years?” The correlation generator allows the user to ask this question by specifying the maximum number of days for which to include downloads and citations after the paper is deposited. We will treat a 30 day period as a ‘month’, and 730 days as 2 ‘years’.

The graphs in [Figure 8.12](#) show downloads (green) and citations (red) that would be included in calculating the correlation for two papers, where downloads were included up to 4 months after deposit and citations up to 2 years. Note: while these papers were deposited in April 2001 (left) and December 1999 (right), some papers that cite them were deposited earlier and may have been citation-linked using a journal citation or may have been updated to include new citations. The citation latency is taken as the time between the first deposit in arXiv of both papers (*i.e.* updates have no bearing on the citation latency, even though author-updates are often done so as to add new citations as indicated by arXiv’s “comment” field).



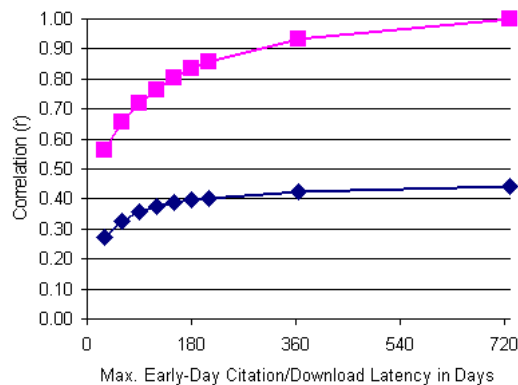
**Figure 8.12:** Example download and citation windows used for prediction calculations.

Given that citation and download impact for the HEP sub-arXiv has a correlation of about  $r = 0.44$  ([Figure 8.11](#)), how long do downloads need to be counted to get

close to this correlation? To test this, queries were made to the correlation generator using 9 different time periods for download data: 1, 2, 3, 4, 5, 6, 7, 12 and 24 months following the deposit of a paper. The correlation increases from 0.270 with one month's of downloads after deposit to 0.440 at 24 months (Table 9.5).

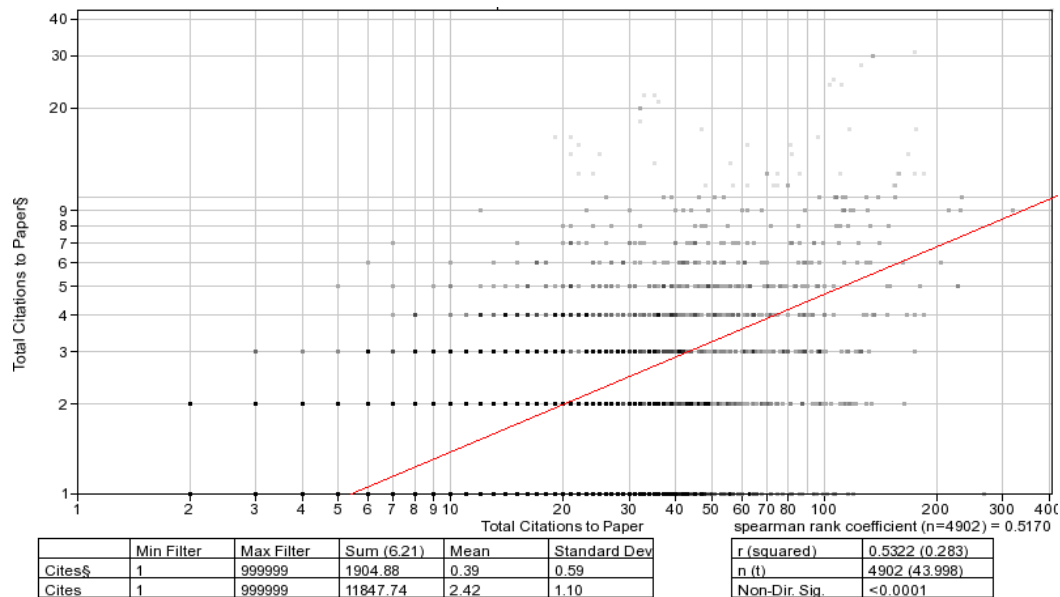
Download/citation correlation (diamond) and hence the power of download counts to predict citation counts reaches an asymptote about 6 months after deposit (Figure 8.13). Download impact at 6 months can predict citation impact at 2 years. For comparison the citation/citation correlation is also plotted (square), showing a higher predictive correlation, possible due to arXiv's rapid distribution model allowing for citations to appear very soon after a paper is deposited. At two years citation/citation correlation is measuring the same values, hence  $r = 1$ .

Figure 8.13 is interesting because it reveals this increase in predictive power is not linear and approximates the correlation found with two years of download and citation data using only 6-7 months of download data. This suggests that if the baseline correlation for a field is significant and sufficiently large, the download data found after 6 months could be used as a predictor of citation impact after 2 years.



**Figure 8.13:** The predictive power of downloads reaches an asymptote at 6 months.





**Figure 8.14:** Correlation between papers' citation impact at one month and two years.

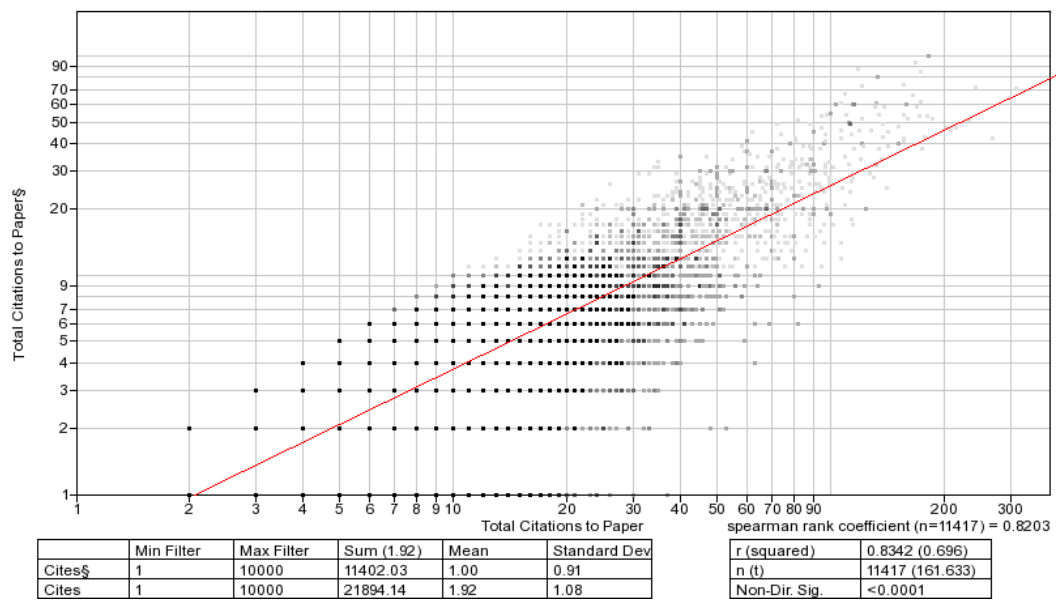
### 8.13 Predicting citation impact from early-day citations

Chapter 10 shows that the peak 'citation latency' for papers deposited in arXiv has decreased from 12 months in the early 1990s to no delay for new papers. The rapid distribution system of arXiv allows for the citation impact to be identified at the earlier pre-print stage in the paper's lifecycle (draft, pre-print, post-print, published *etc.*). Citations to the pre-print within arXiv can be identified and tracked as soon as the paper appears, and those citations could be a predictor of future citation impact.

Comparing one month of citation data to the citation impact at two years for the same 2 year HEP selection of papers results in a correlation of  $r = .5322$

(Figure 8.14). After six months this correlation rises to 0.834 (Figure 8.15).

When compared against the ability of download/citation correlation to predict future citation impact (Figure 8.13) it is apparent that early-day citations provide a stronger base-line correlation, but take longer (hence more citations) to reach the asymptotic point.



**Figure 8.15:** Correlation between papers' citation impact at six months and two years.

## 8.14 Conclusion

Whereas the use of citation counts as a measure of research impact is well established, web-based access to the research literature offers a new potential measure of impact: download counts. Counting downloads is useful for at least two reasons:

1. The portion of later citation variance that is correlated with earlier download counts provides an early-days estimate of probable citation impact that can already begin to be tracked the instant a paper is made Open Access and that already attains its maximum predictive power after 6 months.
2. The portion of download variance that is uncorrelated with citation counts provides a second, partly independent estimate of the usage impact of a paper, sensitive to another research performance indicator that is not reflected in citations ([Kurtz, 2004](#)).

This study found a significant and sizeable correlation between the citation and download impact of papers in physics ( $r = 0.462$ ), as well as in other arXiv fields: mathematics ( $r = 0.347$ ), astro-physics ( $r = 0.477$ ) and condensed matter ( $r = 0.330$ ). This was based only on web downloads from the UK arXiv mirror and those citations that could be automatically found and linked by Citebase. The true correlation may in fact prove higher once more download sites are monitored and automatic linking becomes more accurate: download data are skewed by country (UK-centric due to the mirror location). Because the citation data is only those citations that are ‘internal’ to arXiv (arXiv papers citing other arXiv papers) it will tend to correlate better with arXiv downloads, as presumably arXiv authors are much the same set as arXiv users. The correlation will no doubt vary from field to field, and may also change as the proportion of open access refereed research content (now only 10-20%) approaches 100% (and includes not only refereed journal and conference paper citations and consultations, but books and research data too).

# Chapter 9

## The Effect of Open Access on Citation Behaviour

### 9.1 Introduction

Open access – free, instant access to research findings on the web – has inevitably had an effect on scholarly communication. The strongest evidence for this effect is in the number of downloads that open access papers receive compared to papers available only through subscription: [Kurtz \(2004\)](#) found open access papers received double the number of downloads. Similarly [Richardson \(2005\)](#) found usage of Nucleic Acids Research journal papers doubled when the journal changed to an immediate, free access model. ([Lawrence, 2001](#)) found that the citation impact of research – the most widespread measure of research impact – increases with open access, although the causes of this increase are less clear. ([Kurtz et al., 2005](#)) summarised three possible causes:

1. *Because the access to the [papers] is unrestricted by any payment mechanism authors are able to read them more easily, and thus they cite them more frequently; the Open Access (OA) postulate.*
2. *Because the [paper] appears sooner it gains both primacy and additional time in press, and is thus cited more; the Early Access (EA) postulate.*

3. *Authors preferentially tend to promote (in this case by posting to the internet) the most important, and thus most citable, [papers]; the Self-selection Bias (SB) postulate.*

This chapter investigates the effect of open access on citation impact and citation latency (the time lag between a paper being published and later cited) and whether the obsolescence of papers (as measured by citations) has changed with the increase in web-based access.

## 9.2 Increased Citation Impact due to Open Access

While the majority of new peer-reviewed research is made available on the web, most can only be accessed by users who have a subscription to that content (usually through an institutional site-licence). Open access to research literature maximises the accessibility of that material by providing a version of that literature available on the web accessible to anyone with an Internet connection. Open access increases the use of research papers by 1) providing access to users that do not have a subscription, 2) increasing awareness by being indexed in open access services and 3) by providing faster and earlier access to research results.

When papers can only be accessed through subscriptions, limited budgets determine that only that material which is deemed to be most useful will be subscribed to. Yet this runs contrary to the spirit of scholarly research; that is to research as widely as possible, and increasingly looking beyond the researcher's own subject and into 'inter-disciplinary' topics. The efficient discovery and re-use of ideas demands the greatest possible availability and breadth of ideas to search within. Open access provides both for the capability to access material, once found, but also to allow inter-disciplinary tools to index and autonomously discover related material in disparate fields of enquiry.

While the goal of open access is to maximise accessibility, to motivate authors to provide open access requires demonstrating the benefits to authors (*i.e.* without a clear self-interest in open access, there are plenty of competing demands on authors' time that will prevent open access ever happening). As authors are

increasingly evaluated by their institutions and funders, so the primary record of an author's work (their publications) become ever more important. The simple 'Publish or Perish' culture of counting publications has given way to counting citations (a measure of research 'impact'), either through the proxy of the journal the paper is published in or, as the tools become available, evaluating the author's papers directly.

Clearly if an author is evaluated on the basis of the number of citations to their work it is in their interest to maximise the number of citations they receive. Logically open access should increase citation impact: the greater the number of users, the greater the chance authors will be amongst those users and hence will cite the author. This is especially important with the increasing dominance of general web tools – Google is now as indispensable a research tool as any A&I service *e.g.* for the *Nucleic Acids Research* journal referrals from web search engines went from 1% in 2003 to nearly half in 2006 ([Richardson, 2006](#)). Such generic tools can only index what's on the web, so a user whose work isn't indexed by and easily accessible from a Google link is hence invisible to many potential users.

### 9.2.1 Methodology

To test whether open access (through author self-archiving) increases citation impact a study was performed that coupled the citation data from the Institute for Scientific Information (ISI) CD-ROM citation database, which covers the main journals in medicine, science and technology (see [subsection 3.2.1](#)), and from the arXiv where authors can self-archive their papers to make them open access.

References in the ISI citation database didn't include an identifier for the cited paper, but were processed into bibliographic fields and normalised (journal abbreviation, volume, year *etc.*). Therefore each reference was linked to the cited paper by matching the journal, volume, year and page (of 317 million references 160 million were linked). Author self-citations were identified by flagging citations where the cited and citing paper shared a common author's name (this is ambiguous, as many authors may share the same name, leading to false positives). The number of citations per cited paper was calculated to find the citation impact.

To determine whether a paper in the ISI database was also ‘open access’ (*i.e.* has a version in arXiv) required comparing the bibliographic records from the ISI database against the bibliographic records from arXiv. To find matches the first author and a normalised title were used from ISI and arXiv and matched together. Of course, if the first author were different, or the title substantially different, no match would be found and the ISI paper wouldn’t be flagged as being open access. The citation impact of papers flagged as ‘OA’ and those unflagged could then be compared.

A similar study by [Antelman \(2004\)](#) reports on a comparison of a hand-sample of open and non-open access papers using citation data from ISI. OA papers were determined by human-querying of the Google web search engine. Antelman found a difference of between 45% (Philosophy) and 91% (Mathematics) between the mean citation impact of non-OA and OA papers (*i.e.* 45% more citations).

### 9.2.2 Results

The number of times each paper is cited was calculated for all physics and mathematics paper indexed by ISI from 1992-2003. These papers were divided into those with a version in arXiv and those without. Papers with an arXiv equivalent were defined as being open access (OA). The ‘Non-OA’ papers may have also been available for-free elsewhere on the web, which is part of another on-going study (see [Hajjem et al., 2005](#)) that uses an automated web robot.

The OA advantage is calculated by finding the ratio of the citation counts for OA *vs.* non-OA papers:

$$OAA = 100 \frac{OA}{NonOA} - 100 \quad (9.1)$$

where  $OAA$  is the open access advantage (a percentage),  $OA$  is the average citations to open access papers and  $NonOA$  is the average citations to non-open access papers.

Virtually all of the OA impact effect is positive: OA enhances citation impact substantially *e.g.* between 2-2.5 times as many citations for the *Physics* subject ([Figure 9.6](#)). It would be counter-intuitive to see lower (on average) citation

|                            |       |               |      |            |                          |
|----------------------------|-------|---------------|------|------------|--------------------------|
| <b>Physics:</b>            | 10.1% | 106040/930059 | 134% | 13.95/6.16 | <input type="checkbox"/> |
| Acoustics                  | <1%   | 15/18797      | 109% | 3.97/2.27  | <input type="checkbox"/> |
| Applied Physics            | <1%   | 1970/245265   | 60%  | 7.85/5.73  | <input type="checkbox"/> |
| Chemical Physics           | 1.1%  | 1142/104175   | 49%  | 12.47/9.26 | <input type="checkbox"/> |
| Fluids & Plasmas           | 3.6%  | 845/22305     | 95%  | 13.39/6.01 | <input type="checkbox"/> |
| General Physics            | 13.8% | 43886/267141  | 153% | 15.16/6.14 | <input type="checkbox"/> |
| Miscellaneous Physics      | 16.5% | 1021/4737     | 20%  | 6.42/5.76  | <input type="checkbox"/> |
| Nuclear & Particle Physics | 38.6% | 44798/68470   | 120% | 14.07/6.53 | <input type="checkbox"/> |

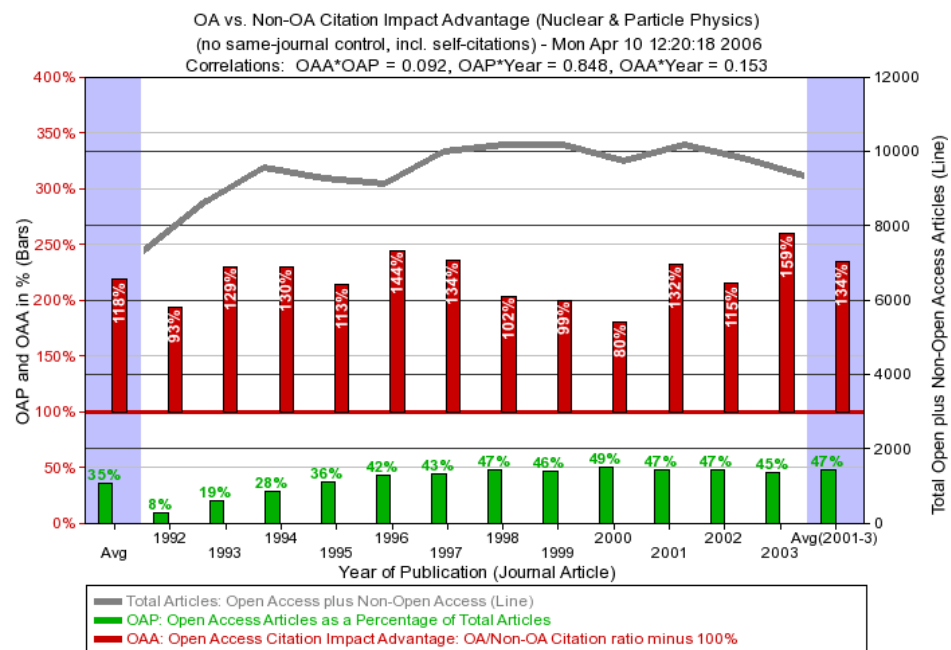
**Figure 9.1:** Subject selection form.

impact for OA papers: all of the papers studied are from the ISI database (hence published in a journal), the arXiv version is *in addition to* the published journal version.

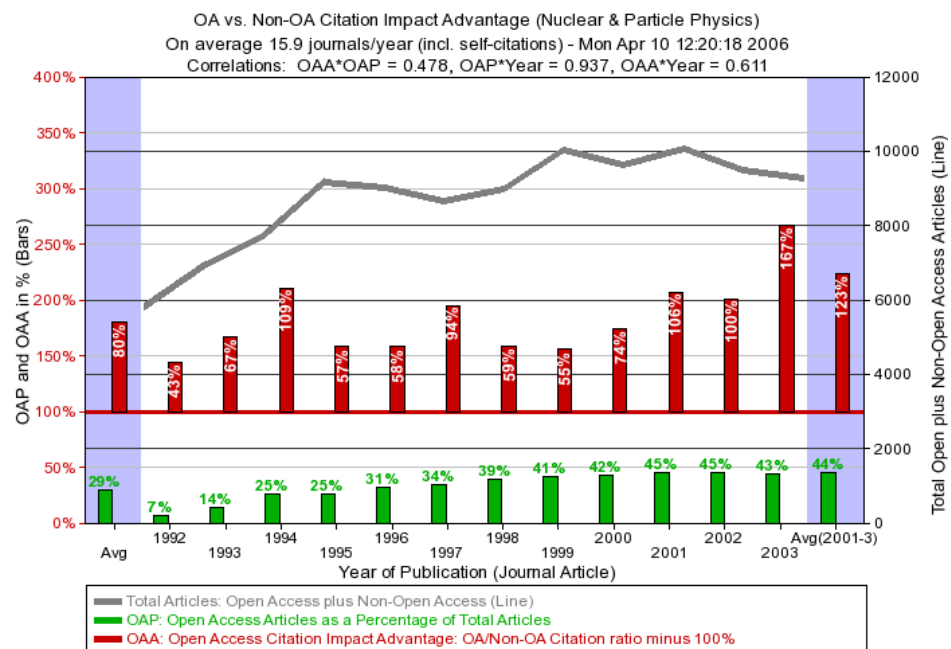
Another measure of interest is the percentage of the yearly papers in a field that are self-archived and thereby made OA relative to the total number of papers in the field. This percentage is slowly increasing across the years and is already especially high in nuclear and high-energy physics, the fields in which arXiv began.

A simple tool was created that allows ISI subject(s) to be selected for analysis. The selection form (**Figure 9.1**) shows all of the available ISI subjects, with shading indicating the percentage of papers that are OA (darker shading of green = more OA), as well as figures for the OA advantage. Each ISI subject area (*e.g.* Physics) and field (*e.g.* Nuclear & Particle Physics) has a tick-box. Ticking a box and clicking ‘Show’ renders 4 graphs for each selected field *e.g.* figures **9.2**, **9.3**, **9.4** and **9.5**.

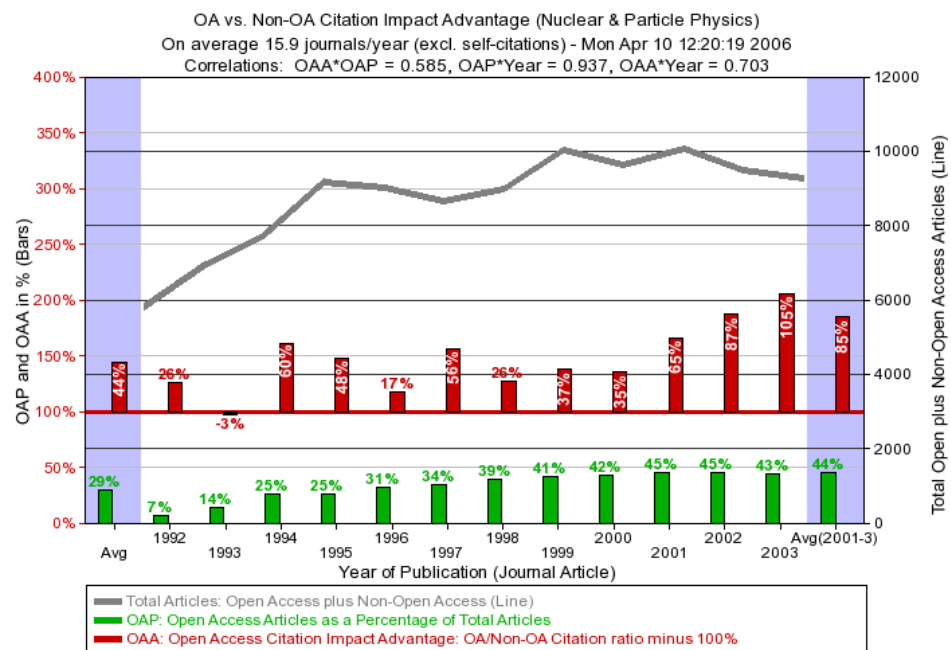




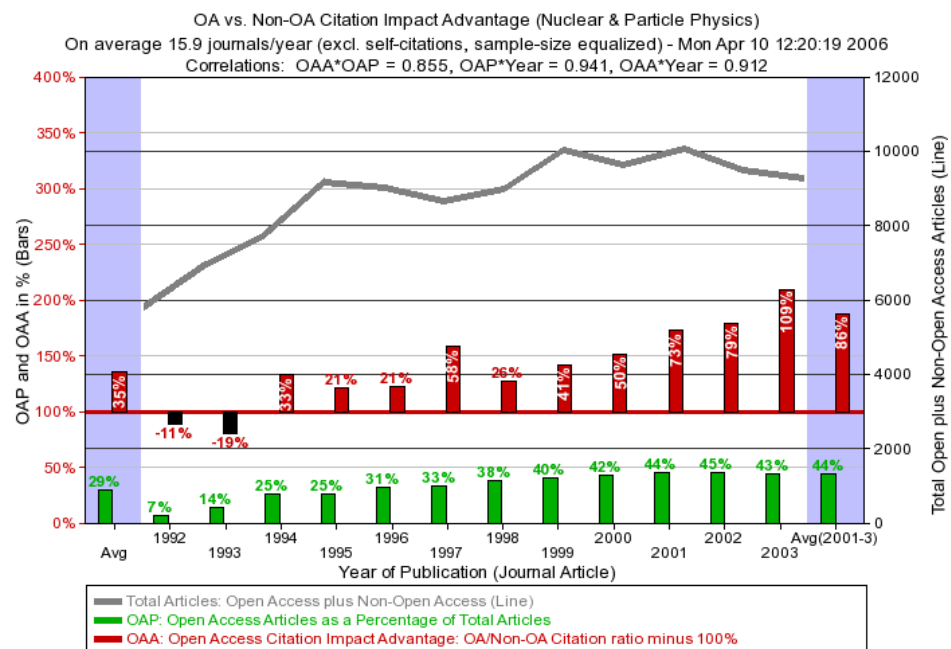
**Figure 9.2:** Open access advantage for Nuclear & Particle Physics, controlled by by-subject



**Figure 9.3:** Open access advantage for Nuclear & Particle Physics, controlled by by-journal



**Figure 9.4:** Open access advantage for Nuclear & Particle Physics, controlled by by-journal and excluding self-citations



**Figure 9.5:** Open access advantage for Nuclear & Particle Physics, controlled by by-journal, excluding self-citations and using same-size samples

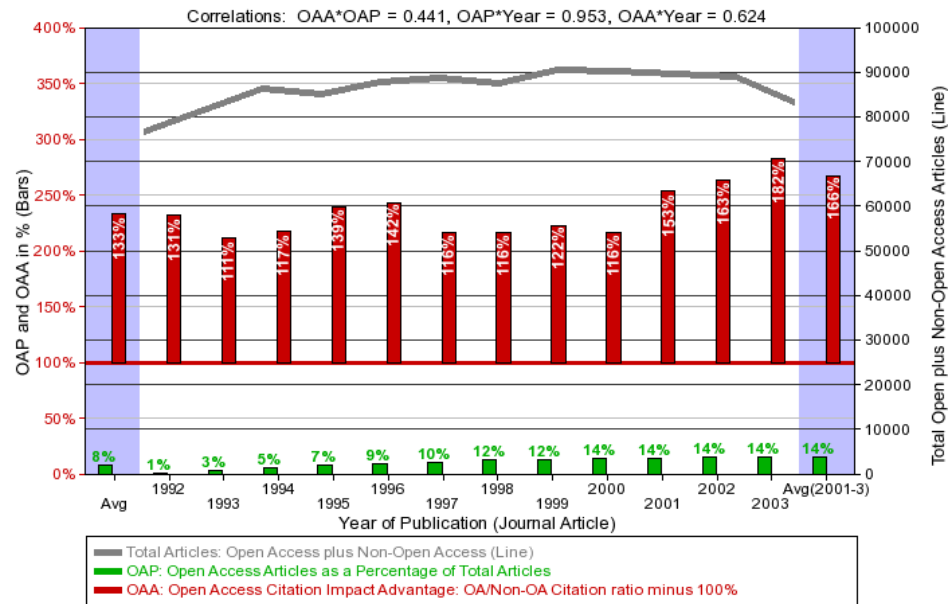
The first graph (*e.g.* [Figure 9.2](#)) is the raw averages: all OA papers in the field compared to all non-OA papers, by year. The second graph (*e.g.* [Figure 9.3](#)) is a more sensitive, controlled comparison, comparing OA *vs.* non-OA papers only within the same journal, by year. The third graph (*e.g.* [Figure 9.4](#)) excludes all author self-citations. The fourth graph (*e.g.* [Figure 9.5](#)) is based on a random sample of non-OA papers within the same journal, rather than all of them (to make the numbers more equal, because there are far more non-OA papers in most journals than OA ones). As each subsequent graph has a smaller  $n$ , it is subject to more noise, hence the OA advantage reduces (as citation data is highly skewed it is particularly susceptible to noise).

The first and last entry in each graph is (first) the average across all 12 years and (last) the average for the most recent three years (2001-2003). The correlations between the OA proportion and time (the OA proportion is growing across the years), between the OA advantage and time and between the OA proportion and the OA advantage are shown. The OA advantage seems to be the strongest within the first 3 years of publication (including a year before publication for the preprint).

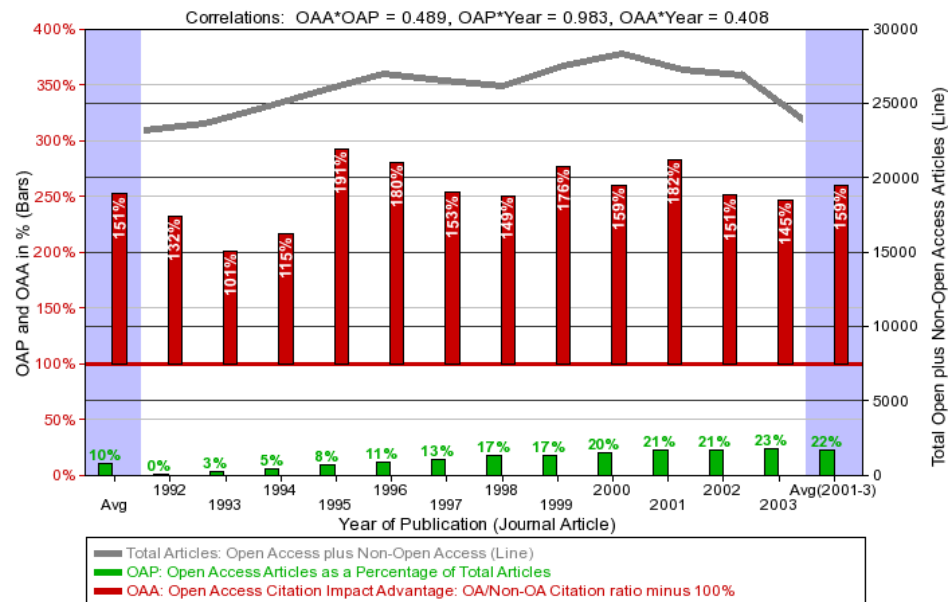
The line represents the total papers counted from the ISI database. The central, red bars represent the OA advantage (*i.e.* 0% = equal citations to OA and Non-OA papers, 100% = twice as many citations to OA *vs.* Non-OA, *etc.*). The lower, green bars are the percentage of the total literature also available in the arXiv.

For the ‘Physics’ field – as defined by ISI – from 1% to 14% of all the included papers were found in arXiv ([Figure 9.6](#)). These OA papers received between 111% and 182% more citations for papers compared to journal-only papers.

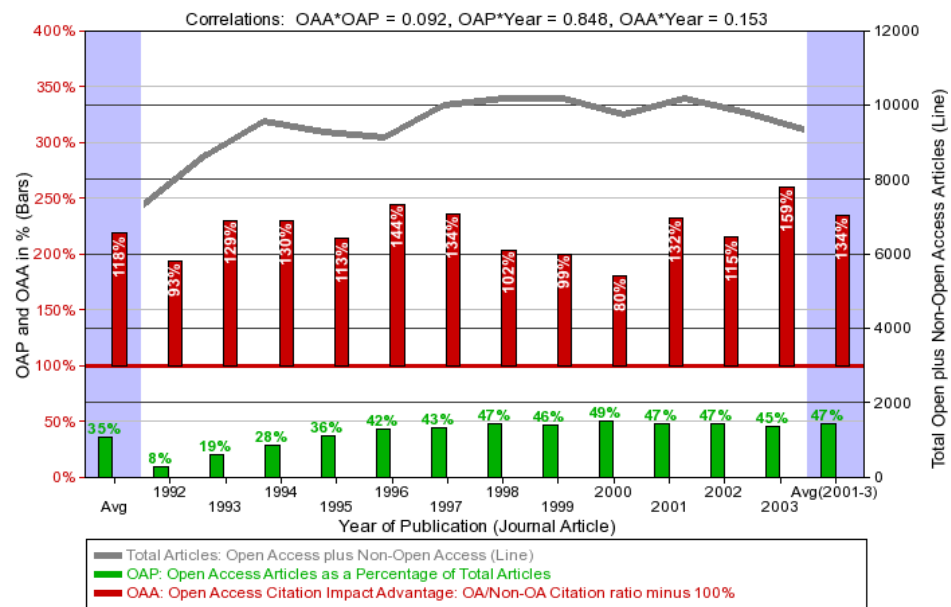
General Physics ([Figure 9.7](#)) is a sub-area of ‘Physics’ and showed a similar percentage of OA papers and OA advantages. Unlike physics this subject did not have a pronounced ‘early days’ advantage in the 3 year period post-publication. Nuclear & Particle Physics most closely fits the subjects covered by the arXiv, hence having up to 45% of a year’s papers also in the arXiv. Despite the high percentage of material available as OA, the OA advantage appears to be smaller (suggesting that as a field approaches saturation the OA advantage will diminish to nothing, as all the citing activity will occur within the OA subset).



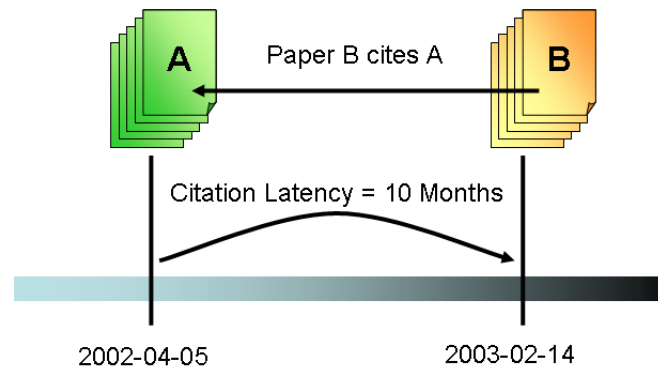
**Figure 9.6:** Open access advantage for all Physics fields.  $OAA*OAP$   $r = 0.441$ ,  $OAP*Year$   $r = 0.953$ ,  $OAA*Year$   $r = 0.624$



**Figure 9.7:** Open access advantage for General Physics.  $OAA*OAP$   $r = 0.489$ ,  $OAP*Year$   $r = 0.983$ ,  $OAA*Year$   $r = 0.408$



**Figure 9.8:** Open access advantage for Nuclear & Particle Physics.  $OAA*OAP$   $r = 0.092$ ,  $OAP*Year$   $r = 0.848$ ,  $OAA*Year$   $r = 0.153$



**Figure 9.9:** Citation latency is the time delay between a paper *A* being deposited and a citing paper *B* being deposited

### 9.3 Reduced Citation Latency due to Pre-Print Archiving

The duration of the Write-Read-Cite cycle is a possible measure for the efficiency of scholarly communication. An author writes a paper (*A*), is read by another author and is then cited by that other author in another paper (*B*). The delay between *A* and *B* is the amount of time it takes for paper *A* to be published (as in to be made public), read, and built upon by other authors – in essence its citation latency.

The time between a citing and cited paper being published is defined as ‘citation latency’. In the example in [Figure 9.9](#) a paper *A* is published on 2002-04-05. A subsequent paper *B* is published on 2003-02-14 and cites the previous paper. The time difference between these two dates is 10 months, hence the ‘citation latency’ for this pair of papers is 315 days. Because arXiv date-stamps papers to the nearest day (and – being the web – papers are instantly accessible) it is possible to use an accuracy of days; something not possible with printed journals.

Citation latency is measure of the efficiency of research communication. While research may be citable for a very long time – *e.g.* in Geology, where the properties and rules of the natural world don’t change – most citations tend towards recent research (as will be shown). As the volume and pace of research inexorably increases, so research publication must follow suit. It is no longer tenable to have research papers languishing in the publication cycle for years when scientific understanding can change dramatically in months. The web – and

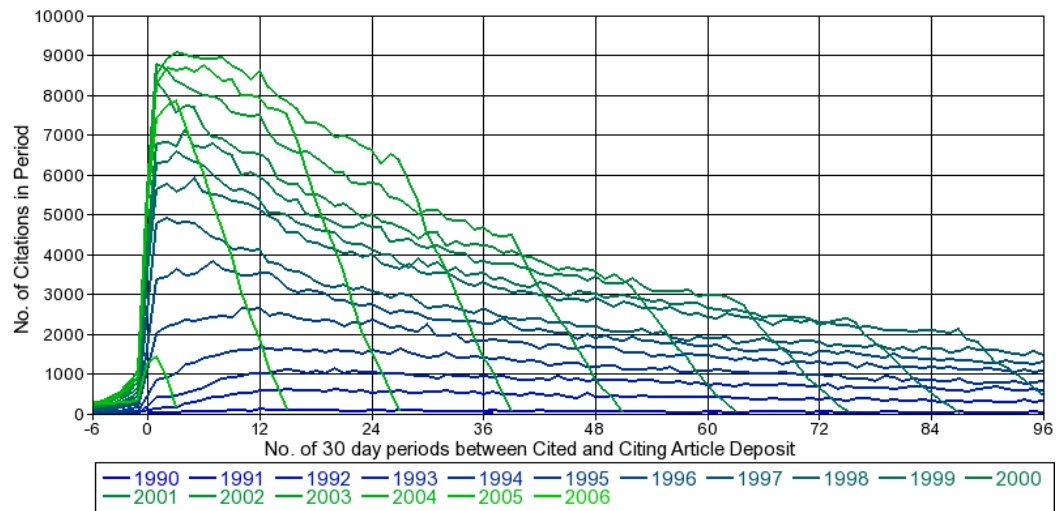
pre-printing in particular – has a profound effect on the speed with which research results can be made public. Modern communications technology allows probes on a glacier in Northern Europe to instantly transmit observation data to a UK-based server (Martinez et al., 2004), and yet the speed with which the results of that research is published can be positively glacial. Rapid pre-printing in the arXiv demonstrates the effect that instant access to research results can have – using the arXiv as a case-study it is possible to see this effect in action over its 15 years.

### 9.3.1 Decreasing e-print Citation Latency

To analyse the effect that the arXiv has had on physics communication the citation latency has been plotted for each year (Figure 9.10). All of the citations for each year are plotted according to the number of days between the citing and cited paper being posted to the arXiv. As the arXiv has been growing linearly since its inception the number of citing papers (hence citations) has grown linearly year on year. This is reflected in older years having a lower line. The oldest (hence lowest) years show a steady increase in the number of citations to a peak at around 12 months, then decreasing over time. The location of this peak-point of citations has shifted each year until the most recent years where there is no apparent delay. This suggests that as more physicists have deposited their papers in the arXiv so they have also increasingly cited arXiv papers and, with the near-instant distribution nature of the arXiv, so the peak age of cited papers has reduced.

It should be kept in mind that the citation data used are from Citebase, hence are only citations to those papers that can be located in the arXiv. This data set tends to emphasise citations to the pre-print (where an author has cited the arXiv identifier) as those citations are most easily identified by Citebase (Figure 7.4, page 100, shows the number of citations included in this analysis *vs.* the total number of citations made by authors).

The advantage of a reducing citation latency is that as the time it takes for new ideas to be published, read, and built upon decreases, so the efficiency of scholarly communication increases, and consequently the efficiency of scholarly research. This is the invisible result of rapid pre-printing within the physics



**Figure 9.10:** Annual citation latency for arXiv

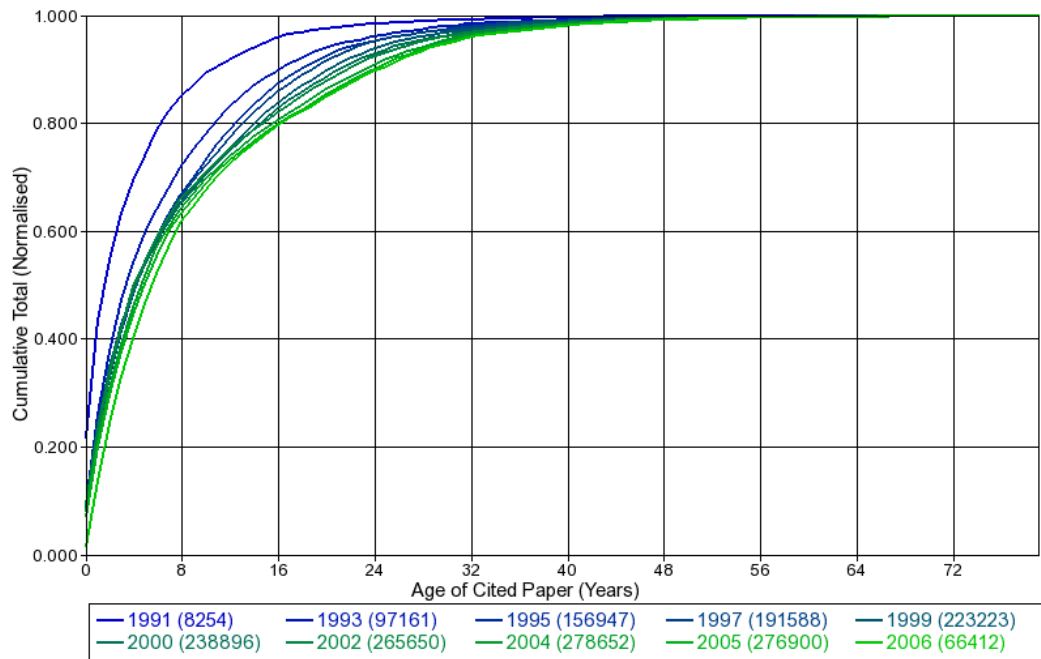
community. Open access also provides other unseen benefits, in simply reducing the amount of time taken by researchers to locate relevant material. arXiv, and the research tools that harvest it, provide faster ways for researchers to discover and access scholarly literature.

## 9.4 Citation Obsolescence

While Citebase’s citation database allows for a fine-grained analysis of citation latencies it is possible to look at the temporal properties of all references, but only with a per-year granularity; every citation to a journal paper includes the publication year of the journal (although that may not be the actual year a journal was actually published, it will be accurate to at least one year). Each year of papers deposited in the arXiv can have their references plotted against the age of the cited paper. This provides an insight into the currency of research in that field – or, to look at it another way, how long a paper might expect to continue to be cited by new papers.

Figure 9.11 shows the age of papers cited each year from arXiv papers for the High Energy Physics field (the most complete arXiv sub-field); the average age of the cited literature has actually increased – from *e.g.* in 1993 eighty percent of all citations were to papers 10 years or younger while more recently that has grown to 16 years. A possible reason for this could be that with the advent of electronic

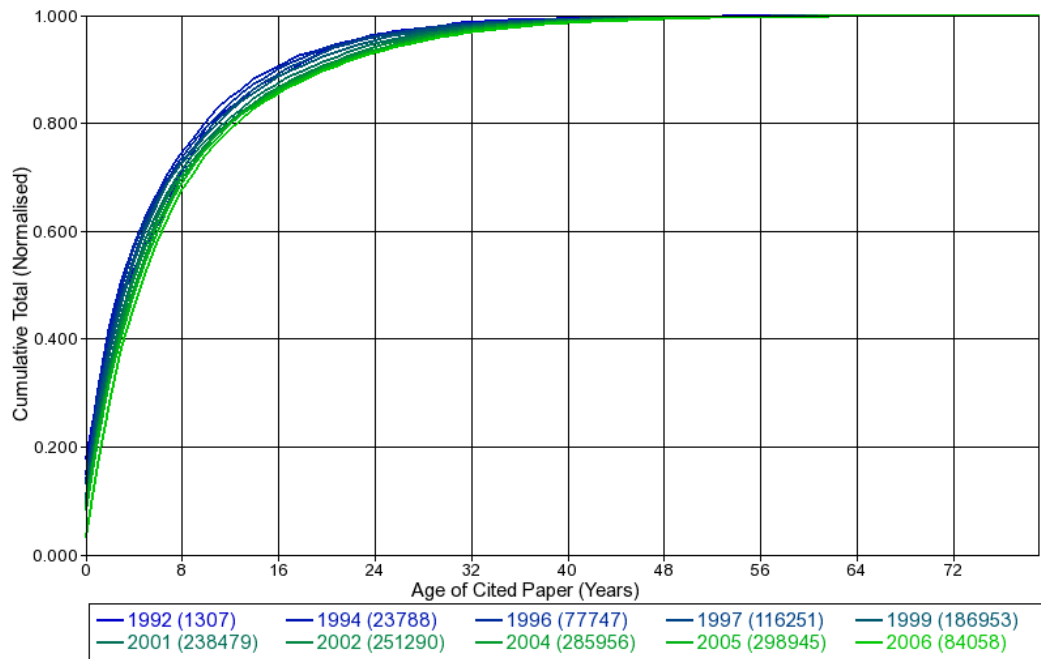




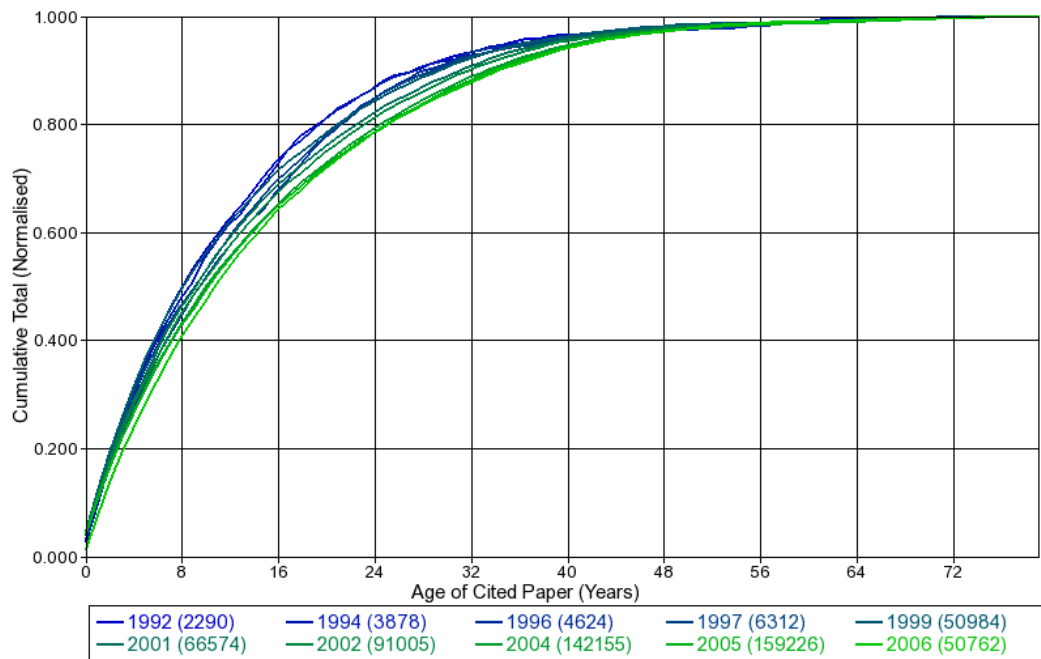
**Figure 9.11:** The age of papers cited by arXiv High Energy Physics papers

access it becomes as easy to access a 50-year paper as it is a paper published today (compared to locating journal papers in a library's archive). However, other fields appear to show different changes over time, *e.g.* in Astrophysics (Figure 9.12) there is considerably less variation across years. Maths (Figure 9.13) shows a similar increase in the age of cited papers, but for all years the age of cited papers is much higher than for physics.

Having a record of the age of cited literature gives an idea of how much back-catalogue is needed to have a reasonable chance of being able to locate a cited paper. The number of monthly deposits to the High Energy Physics (HEP) arXiv sub-field has remained almost constant in the 11 years to 2005 (see Figure 8.2, section 8.3), suggesting that for those 11 years most of the published HEP literature has also been deposited in arXiv by authors. Table 9.1 shows what percentage of cited papers would be found given increasing numbers of existing years included. For HEP 11 years covers approximately 71% of cited papers, or 10 'missed citations' (given the long-tail nature of the graph, 22 years gets another 15% *etc.*). Reducing the number of missed citations to just 2 would require a total of 30 years of existing records.



**Figure 9.12:** The age of papers cited by arXiv Astronomy/Astrophysics papers



**Figure 9.13:** The age of papers cited by arXiv Maths papers

**Table 9.1:** Cumulative percentage of cited literature by age for High Energy Physics e-prints.

| Years Included | % Cited Refs | Missed Citations |
|----------------|--------------|------------------|
| 0              | 6.55%        | 29               |
| 1              | 19.64%       | 25               |
| 2              | 29.53%       | 22               |
| 3              | 37.58%       | 19               |
| 4              | 44.43%       | 17               |
| 5              | 50.18%       | 16               |
| 10             | 68.68%       | 10               |
| 20             | 85.02%       | 5                |
| 30             | 94.70%       | 2                |
| 40             | 97.91%       | 1                |
| 80             | 100.00%      | 0                |

## 9.5 Conclusion

Open access papers have greater citation impact than non-open access papers: arXiv papers received – on average – twice the number of citations than non-arXiv papers. The peak citation latency for arXiv papers has also reduced to virtually zero (*i.e.* the greatest rate of citations to arXiv papers, by arXiv papers, is straight after being deposited). At the same time, the total age of cited papers has increased, particularly in the High Energy Physics (HEP) arXiv sub-field.

These results suggest that open access to research papers provides a citation impact advantage (at least in the transitional stage where only part of the literature is open access) and reduces the citation latency (because papers are accessible sooner). Because electronic access makes older papers more accessible so they will be read (and cited) more, hence highly-cited open access papers will not only be cited sooner (and more) but also for longer.

The number of ‘missing citations’ (as discussed in the previous section) is both an open access and digital preservation issue. As citation-based services are built on open access papers, their usefulness depends on the facilitating access to older papers, as the user follows citations. Studying the age of cited papers provides a measurable amount of backlog necessary to cover a certain percentage of the cited literature. For digital preservation of research literature it could be argued that preserving access to both the *cited* as well as current literature is necessary (otherwise there is no context to current research). Again, measuring the age of

the cited literature provides a quantitative measure of the historical literature necessary to provide context to current research.

# Chapter 10

## Conclusions and Future Directions

### 10.1 Introduction

The Serials Crisis has helped to highlight a greater problem in the scholarly communication system: because the predominant economic model of scholarly journals is based on user-pays, subscription-based access so many potential users are denied access. This loss of access not only effects the ability of research users to perform their research but also (because those researchers may also be authors themselves) reduces the impact of authors by denying them potential users, hence potential citations.

The solution to the access and impact crisis is ‘open access’ – providing free access to the scholarly material through either a parallel system of ‘author self-archiving’ or by moving towards a sponsored or author-pays type model of publication. This is made possible by the essentially free cost of posting information on the web (because researchers’ institutions have already paid for the web infrastructure, making the incremental storage and bandwidth costs negligible).

To facilitate and analyse the state of open access I have developed several services. Celestial harvests metadata from subject and institutional based repositories of author self-archived research papers. In conjunction with the Registry of Open Access Repositories I can analyse the number of new

institutional repositories and the amount of content they contain. Citebase Search goes a step further by harvesting the full-text e-prints from a number of repositories, extracts and links their references and builds a citation index. This citation index has allowed me to create an open access citation navigation and analysis tool, as well as being able to look more closely at the effects of open access on the citation behaviour of authors. I have used web usage logs as a new metric for author impact – download impact – and tested how good a predictor web usage is of citation impact.

## 10.2 Contribution of this Thesis

*The main contribution of this work is to demonstrate the effectiveness of open access as a means of scholarly communication.*

This thesis represents work undertaken in an uncertain and emerging field. If there is no clear central hypothesis it is because this work has been developed and adapted to changing circumstances and demands. Therefore the main contribution of this thesis is somewhat *avant-garde*: that I have been able to create tools and perform bibliometric research with only my own time and minimal technical support from my host institution demonstrates the benefit of open access (compared to having to ‘buy-in’ large databases from *e.g.* ISI).

More concrete contributions are the discovery of the reducing citation latency in arXiv citations and the difference in citation impact of arXiv papers compared to other papers in the same subject. This work has also contributed to the design of the OAI-PMH and demonstrated the need to monitor and aggregate institutional repository collections. The Citebase, Celestial and ROAR services remain as valuable tools for use by users, bibliometricians and the digital library community.

Parts of the programming code used in this thesis have been released as open source for re-use by others: the OAI-PMH library (that abstracts across OAI-PMH protocol versions, handles errors *etc.*), the Celestial OAI caching tool, the `TeX::Encode` and `Logfile::EPrints` Perl modules (part of Citebase) and contributions to the Perl `XML::LibXML` (threading-support fixes) and `Xapian::Search` modules (QueryParser wrapper and other fixes).

### 10.3 Open Access Improves Citation Impact and Decreases Citation Latency

To test whether open access increases citation impact I used the arXiv – a collection of author self-archived physics, maths and computer-science eprints. Comparing the number of citations to journal papers with and without an eprint in arXiv I have found that papers with an arXiv eprint receive about twice as many citations as not. This is due to several inter-related factors. Foremost arXiv is a rapid distribution mechanism – usually just a single day between an author posting an eprint and it being available for download. This early-days advantage provides additional readers that results in earlier citations. Because papers that are highly cited are read more, this early-day citation advantage persists for the life of the paper (the more links – citations and web hyperlinks – there are to a paper, the higher likelihood it will be seen, read and cited).

Because arXiv is free to access anyone with an Internet connection can access the papers' full-text. arXiv is also extensively used by physics and mathematical researchers, as well as being indexed in Google and other web-tools. This ease of access increases the citation impact of arXiv papers by allowing more people access to the same paper (compared to a subscription-only journal) and making it more likely that active researchers (hence more potential citing authors) will see the paper.

I have found the peak 'citation latency' in arXiv has reduced from upwards of a year to near-instantaneous. As the time it takes for new ideas to be published, read, and built upon has decreased, so the efficiency of scholarly communication increases, and consequently the efficiency of scholarly research. This is the invisible result of rapid pre-printing within the physics community. Open access also provides other unseen benefits, in simply reducing the amount of time taken by researchers to locate relevant material.

## 10.4 Web Impact as a Predictor of Citation Impact

This thesis has looked at the relationship between online download impact (number of downloads of a paper) and citation impact (citations from other papers), and found a reasonable correlation (about  $r = 0.5$ ), despite having only a small proportion of the download data<sup>1</sup>.

Whereas the use of citation counts as a measure of research impact is well established, web-based access to the research literature offers a new potential measure of impact: download counts. Counting downloads is useful for at least two reasons, firstly the portion of later citation variance that is correlated with earlier download counts provides an ‘early-days’ estimate of probable citation impact that can already begin to be tracked the instant an article is made open access. This degree of co-variance reaches its maximum predictive power after six months (*i.e.* the highest correlation starts at six months of download data predicting two years of citation data). Secondly the portion of download variance that is uncorrelated with citation counts provides a second, partly independent estimate of the usage impact of an article, sensitive to another research performance indicator that is not reflected in citations.

It is likely that download impact is just the first of many new and powerful indicators of research impact and direction that will emerge from an Open Access corpus (Hitchcock (2005) has compiled a list of many studies using and looking at OA material) – indicators that will include co-citation analysis (to and from jointly cited or citing papers and authors), co-download analysis (Bollen et al., 2005), co-text analysis (from boolean word conjunctions to latent semantic indexing and other measures of text similarity patterns and lineage, *e.g.* see Deerwester et al. (1990)), citation-based analogues of Google’s recursive PageRank (Brin and Page, 1998) algorithm weighting cited papers’ (or authors’) citation ranks with the citation weights of the citing papers (authors), hub/authority analysis (papers cited by many papers vs. papers citing many papers, see Kleinberg (1999)), and time-series chronometric analyses. Citation and download counts are just the first two terms in what will be a rich and diverse multiple regression equation predicting and tracking research impact.

---

<sup>1</sup>Download data is only available from the UK-based arXiv mirror <http://uk.arXiv.org/>



## 10.5 Maximising the Benefit of Research Funding

The tax-payer funding of research is a public good: intended for the benefit of the society that pays for it. To derive the maximum benefit from this investment the technology and economics of communicating the results of research should allow any researcher to access the work of all other researchers. And yet, most of the results of research – in the form of research papers and monographs – is accessible only on payment of a subscription or other access-charge. The result of these charges is that no researcher can afford access to all research results, so researchers (or more commonly librarians on their behalf) have had to ‘cherry-pick’ the best research results to purchase.

This cherry-picking decreases the benefit that society derives from its investment in research in two ways, firstly through the wasted administrative cost of charging researchers access to the results of research that has already been paid for and secondly, and most importantly, because the reduction in access means researchers potentially miss important and relevant results. What is needed, and now slowly happening, is to remove these inefficiencies.

Charging access to research results is a natural consequence of the costs associated with printing and distributing paper-based documents. By charging per-copy, so income scales with the cost of printing and posting (with hopefully sufficient copies sold to cover the one-time editorial and reviewing costs). However, it is both not feasible and unnecessary to provide every researcher with the results from every piece of research ever published on paper to achieve open access. Instead the solution to removing access-charges lies in the near-zero copy costs of the Internet. Any author can distribute their work in digital form on the Internet for little more than the cost of their own Internet access.

In this thesis I have outlined two, complimentary, methods for providing free (open) access to research results on the Internet. Authors can – in addition to publishing their research in a journal or book – post their work on the Internet through a personal web site, institutional or subject-based repository. Or an author can publish their work in an open access journal, which may charge an up-front fee to cover their administrative costs of editing, reviewing and Internet access but provides its content free to all users. In this way an author can make

their work accessible to potentially an unlimited number of users.

## 10.6 Future Directions

The installation and uptake of institutional repositories is growing rapidly. To analyse institutional repositories requires a service that can aggregate and normalise their content. Because Citebase Search has focused on the arXiv, it has only needed to support the citation styles that arXiv physics and maths authors tend to use (as journal policies in a subject area tend to determine citation style). Extending Citebase to institutional repositories requires support for potentially any research discipline, hence pretty much any citation style. Rather than extending the current mass of rules used to parse references this may require a different technical approach.

As well as supporting a wider variety of subjects, parsing references from digitised works (*e.g.* scanned from printed journals) will grow the volume citation data available. While most digitisation projects include extracting and marking-up references, it is easy to imagine many authors scanning their older papers themselves *e.g.* as Eugene Garfield has done with almost all of his papers. Such papers would end up as PDFs available from institutional repositories, necessitating both OCR and reference extraction/linking if they are to be included in Citebase.

My hope is reference parsing will become unnecessary as authoring tools become more advanced *e.g.* by supporting structured, standardised XML mark-up. Authoring tools could create ‘smart’ documents that understand bibliographic citations, people and other real-world objects. One possible scenerio is that an author ‘copies and pastes’ a special link – an object that contains a bibliographic description of the object along with its URL (or other linking system) – from a web page into their authoring tool. That link is then embedded as structured data (but displayed in the appropriate citation style) that can later be extracted by a citation index for the purposes of citation analysis. The benefit of this approach versus only using a unique-linking scheme (*e.g.* DOI) is that it reduces the potential for mistakes to cause broken or incorrectly targetted links (because there are multiple redudant vectors) and it doesn’t rely on the continuing existence of a central database. Therefore future citations will likely look much as

they do now, but with structured data behind the scenes.

Compared to citations, usage data is relatively easy to collect – web servers already capture what is accessed (and who does the accesses) to log files. To aggregate web logs from multiple sites requires a common mechanism for harvesting those web logs and standardising requests from multiple repository softwares (*i.e.* to ensure ‘full-text’ requests from one repository get counted the same number of times as equivalent requests to another repository). The power of cross-repository usage statistics is that it provides a complete picture for an individual work (by aggregating usage of multiple copies), and allows within-subject comparisons (the collection defined by an institutional repository is limited to the members of that institution, who work in different fields hence aren’t easily comparable).

### 10.6.1 Extending and supporting this work

I support three services for open access: Citebase Search, Celestial and the Registry of Open Access Repositories (ROAR). The best way to ensure continuing support is to make them as indispensable as possible – even if such services are ‘free’, if something is sufficiently useful they can be adopted and supported.

The source code for both Celestial and ROAR can be downloaded, built upon, installed and run by anyone else. By its nature Celestial doesn’t contain any data that isn’t available from the repositories it harvests from. Celestial is hence a convenience, and not necessarily needed. ROAR does contain original data, contributed by institutional repository administrators and edited. My aim is to make ROAR as maintenance-free as possible, by allowing repository administrators to add and maintain their own records. Looking forward, a schema could be created that would allow the ROAR-specific metadata to be defined and exported in the repository’s OAI interface, making the registry just an aggregation tool.

Citebase Search is by far the most heavily used of the three services I’ve created. While I believe the approach I’ve taken (using OAI-PMH) and the analyses I’ve created from Citebase are novel, the core function of building a citation index is implemented by a wide variety of services. However, one particularly useful aspect of Citebase is that its data can be made freely available, because it only uses data

provided by authors. I would like to see Citebase continue in this role or, ideally, to be subsumed by a larger index but still being able to provide that citation data to other researchers, to enable others to create novel analyses, user interfaces *etc.*

In addition to continuing to support and expanding the scope of Citebase Search, it would be interesting to extend the current ranking and navigation features. These might be to create text-similarity metrics (“this text is related to that text by content”), secondary author-metrics (*e.g.* rank by author influence) or to combine multiple metrics in search result rank-ordering (*e.g.* rank by funding-council citation impact and then by paper downloads). The goal is to provide more intelligent tools to end-users, in particular to serve the needs of the new user (who needs access to the ‘seminal’ works in a topic) as well as the needs of the experienced user (who is interested in new works and how their subject interfaces with related topics).

## 10.7 In Summary

For now the development in open access is focused on gaining content, through facilitating author-participation and demonstrating the benefits to the ‘leading edge’ of innovators. As open access becomes the norm so the focus will shift to the new potential uses for a large collection of high-quality content. As research is increasingly desktop-based (an astronomer need never look through a telescope herself), so tying together data sets (*e.g.* astronomical observations) and the research derived from them will become more important. Similarly the research papers themselves can become just another data set, from which patterns and derived findings could be achieved.

# Bibliography

ANSI. The OpenURL Framework for Context-Sensitive Services, 2004. URL [http://www.niso.org/standards/standard\\_detail.cfm?std\\_id=783](http://www.niso.org/standards/standard_detail.cfm?std_id=783).

ANSI/NISO Z39.88-2004. 45

Kristin Antelman. Do Open Access Articles Have a Greater Research Impact? *College & Research Libraries News*, 65(5):372–382, 2004. URL <http://eprints.rclis.org/archive/00002309/>. 144

Helen Atkins. The ISI® Web of Science® – Links and Electronic Journals. *D-Lib Magazine*, 5(9), September 1999. URL <http://www.dlib.org/dlib/september99/atkins/09atkins.html>. 33

Helen Atkins, Catherine Lyons, Howard Ratner, Carol Risher, Chris Shillum, David Sidman, and Andrew Stevens. Reference linking with DOIs: A case study. *D-Lib Magazine*, 6(2), February 2000. URL <http://www.dlib.org/dlib/february00/02risher.html>. 47

Steven Bachrach, R. Stephen Berry, Martin Blume, Thomas von Foerster, Alexander Fowler, Paul Ginsparg, Stephen Heller, Neil Kestner, Andrew Odlyzko, Ann Okerson, Ron Wigington, and Anne Moffat. Intellectual property: Who should own scientific papers? *Science*, 281(5382):1459–1460, 1998. URL <http://www.sciencemag.org/cgi/content/full/281/5382/1459>. vi, 1

Kathleen Bauer and Nisa Bakkalbasi. An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11(9), September 2005. URL <http://dlib.org/dlib/september05/bauer/09bauer.html>. 32

Donna Bergmark, Paradee Phempoonpanich, and Shumin Zhao. Scraping the ACM Digital Library. *SIGIR Forum*, 35(2):1–7, 2001. ISSN 0163-5840. URL <http://doi.acm.org/10.1145/511144.511146>. 34

- J. Bollen, H. Van de Sompel, J. Smith, and R. Luce. Toward alternative metrics of journal impact: A comparison of download and citation data. *ArXiv Computer Science e-prints*, March 2005. URL <http://arXiv.org/abs/cs/0503007>. 118, 161
- Christine L. Borgman and Jonathan Furner. Scholarly communication and bibliometrics. In B. Cronin, editor, *Annual Review of Information Science and Technology*, volume 36, pages 3–72. Information Today, 2002. URL <http://polaris.gseis.ucla.edu/jfurner/arist02.pdf>. 117
- Joseph Branin, Frances Groen, and Suzanne Thorin. The changing nature of collection management in research libraries. Technical report, Association of Research Libraries, 1999. URL <http://www.arl.org/collect/changing.html>. 19
- Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998. URL <http://www.cis.upenn.edu/~yicn/cse591f05/anatomy.pdf>. 94, 161
- Tim Brody, Ian Hickman, and Stevan Harnad. Mining the Social Life of an ePrint Archive, 2000. URL <http://opcit.eprints.org/tdb198/opcit/>. Accessed 2006-03-03. 1
- Leslie Carr and Stevan Harnad. Keystroke economy: A study of the time and effort involved in self-archiving. Technical report, University of Southampton, 2005. URL <http://eprints.ecs.soton.ac.uk/10688/>. 22, 77
- Leslie Carr and John MacColl. IRRA (Institutional Repositories and Research Assessment) RAE Software for Institutional Repositories. Technical report, IRRA Project, 2005. URL <http://irra.eprints.org/white/>. White paper. 23, 29
- Leslie Chan and Barbara Kirsop. Open archiving opportunities for developing countries: towards equitable distribution of global knowledge. *Ariadne*, (30), December 2001. URL <http://www.ariadne.ac.uk/issue30/oai-chan/>. 4
- Leslie Chan, Darius Cuplinskas, Michael Eisen, Fred Friend, Yana Genova, Jean-Claude Guédon, Melissa Hagemann, Stevan Harnad, Rick Johnson, Rima Kupryte, Manfredi La Manna, István Rév, Monika Segbert, Sidnei de Souza, Peter Suber, and Jan Velterop. Budapest Open Access Initiative, 2002. URL <http://www.soros.org/openaccess/read.shtml>. 2, 9

- John Cox and Laura Cox. Scholarly publishing practice: the ALPSP report on academic journal publishers policies and practices in online publishing. Technical report, ALPSP, 2003. URL <http://www.alpsp.org/news/sppssummary0603.pdf>. 16
- Raym Crow. The case for institutional repositories: A sparcs position paper. Technical report, The Scholarly Publishing & Academic Resources Coalition, 21 Dupont Circle Washington, DC 20036, 27 August 2002. URL <http://www.arl.org/sparc/IR/ir.html>. 9, 73
- Philip M. Davis and Suzanne A. Cohen. The effect of the Web on undergraduate citation behavior 1996-1999. *Journal of the American Society for Information Science and Technology*, 52(4):309–314, 2001. doi: 10.1002/1532-2890(2000)9999:9999<::AID-ASI1069>3.0.CO;2-P. URL [http://dx.doi.org/10.1002/1532-2890\(2000\)9999:9999<::AID-ASI1069>3.0.CO;2-P](http://dx.doi.org/10.1002/1532-2890(2000)9999:9999<::AID-ASI1069>3.0.CO;2-P). 16
- Michael Day. Institutional repositories and research assessment. Technical report, UKOLN, 2 December 2004. URL <http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/rae/rae-study.pdf>. 23, 29
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990. doi: doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL <http://www.si.umich.edu/~furnas/Papers/LSI.JASIS.paper.pdf>. 161
- Tony Delamothe and Richard Smith. Pubmed central: creating an aladdin’s cave of ideas. *British Medical Journal*, 322:1–2, 6 January 2001. URL <http://bmj.bmjjournals.com/cgi/content/full/322/7277/1>. 13
- Peter J. Denning. The ACM digital library goes live. *Commun. ACM*, 40(7): 28–29, 1997. ISSN 0001-0782. URL <http://doi.acm.org/10.1145/256175.256179>. 34
- A Fassoulaki, A Paraskeva, K Papilas, and G Karabinis. Self-citations in six anaesthesia journals and their significance in determining the impact factor. *Br. J. Anaesth.*, 84(2):266–269, 2000. URL <http://bj.a.oxfordjournals.org/cgi/content/abstract/84/2/266>. 56

Nancy Fried Foster and Susan Gibbons. Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine*, 11(1), January 2005. URL <http://www.dlib.org/dlib/january05/foster/01foster.html>. 22

Eugene Garfield. Citation Indexes for Science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 15 July 1955. URL [http://www.garfield.library.upenn.edu/papers/science\\_v122v3159p108y1955.html](http://www.garfield.library.upenn.edu/papers/science_v122v3159p108y1955.html). 32, 54, 55

Eugene Garfield. Science citation index – a new dimension in indexing. *Science*, 144(3619):649–654, May 1964. URL <http://www.garfield.library.upenn.edu/essays/v7p525y1984.pdf>. 32

Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972. URL <http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf>. 54

Eugene Garfield. How ISI selects journals for coverage: Quantitative and qualitative considerations. *Current Contents*, 28 May 1990. URL <http://www.garfield.library.upenn.edu/essays/v13p185y1990.pdf>. 55

Eugene Garfield. The impact factor. *Current Contents*, 25:3–7, 20 June 1994. URL <http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html>. 117

Eugene Garfield. The agony and the ecstasy - the history and the meaning of the journal impact factor, 2005. URL <http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf>. Presented at the International Congress on Peer Review and Biomedical Publication, Chicago, U.S.A. September 16, 2005. 56

Eugene Garfield. Web Site of Eugene Garfield, Ph.D., 2002. URL <http://www.garfield.library.upenn.edu/>. 54

Eugene Garfield and I.H. Sher. New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3):195–201, July 1963. URL <http://www.garfield.library.upenn.edu/essays/v6p492y1983.pdf>. 54

Aldo Geuna and Ben R. Martin. University Research Evaluation and Funding: an International Comparison. *Minerva*, 41:277–304, 2003. URL



- <http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp71/sewp71.pdf>. 29
- Paul Ginsparg. Los Alamos XXX. *APS News Online*, 1996. URL <http://www.aps.org/apsnews/1196/11718.cfm>. 18
- Paul Ginsparg. *The Transition from Paper: Where Are We Going and How Will We Get There?*, chapter Electronic Clones vs. the Global Research Archive. American Academy of Arts & Sciences, 2001. 14, 18
- Paul Ginsparg. Can peer review be better focused?, 2003. URL <http://arxiv.org/blurp/pg02pr.html>. 118
- Google. Google Scholar Support for Scholarly Publishers, 2005. URL <http://scholar.google.com/scholar/publishers.html>. [Accessed 2006-03-07]. 27
- P.L.K. Gross and E.M. Gross. College libraries and chemical education. *Science*, 66(1713):385–389, 1927. URL <http://links.jstor.org/sici?sici=0036-8075%2819271028%293%3A66%3A1713%3C385%3ACLACE%3E2.0.CO%3B2-9>. 55
- Christopher Gutteridge. GNU EPrints 2 overview. In *Proceedings of the 11th Panhellenic Academic Libraries Conference*, Greece, 2002. URL <http://eprints.ecs.soton.ac.uk/6840/>. 71
- Chawki Hajjem, Stevan Harnad, and Yves Gingras. Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *IEEE Data Engineering Bulletin*, 28(4):39–47, 2005. URL <http://eprints.ecs.soton.ac.uk/11688/>. 5, 15, 18, 144
- Stevan Harnad. *Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing*, chapter A subversive proposal. Association of Research Libraries, Washington, DC, June 1995. URL <http://www.arl.org/scomm/subversive/toc.html>. 72
- Stevan Harnad. The self-archiving initiative: Freeing the refereed research literature online. *Nature*, 410:1024–1025, 26 April 2001a. doi: 10.1038/35074210. URL <http://www.nature.com/nature/debates/e-access/Articles/harnad.html>. 14

- Stevan Harnad. For whom the gate tolls? how and why to free the refereed research literature online through author/institution self-archiving, now, 2001b. URL <http://cogprints.org/1639/>. 21, 72
- Stevan Harnad. Generic rationale and model for university open access self-archiving mandate, March 13 2006. URL <http://openaccess.eprints.org/index.php?/archives/71-guid.html>. 73
- Stevan Harnad and Tim Brody. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, 10(6), 2004. URL <http://www.dlib.org/dlib/june04/harnad/06harnad.html>. 27
- Stevan Harnad, Tim Brody, Francois Vallieres, Les Carr, Steve Hitchcock, Gingras Yves, Oppenheim Charles, Heinrich Stamerjohanns, and Eberhardt Hilf. The access/impact problem and the green and gold roads to open access. *Serials review*, 30(4), 2004. URL <http://eprints.ecs.soton.ac.uk/10209/>. 4, 13, 24, 26, 28, 55
- Steve Hitchcock. Explore open archives: Core metalist of open access eprint archives, 2003. URL <http://opcit.eprints.org/exploresearchives.shtml>. 74
- Steve Hitchcock. The effect of open access and downloads ('hits') on citation impact: a bibliography of studies, 2005. URL <http://opcit.eprints.org/oacitation-biblio.html>. 161
- Steve Hitchcock, Les Carr, Zhuoan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze, and Stevan Harnad. Developing services for open eprint archives: globalisation, integration and the impact of links. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 143–151, June 2000. URL <http://eprints.ecs.soton.ac.uk/2860/>. 1, 6, 38
- Steve Hitchcock, Donna Bergmark, Tim Brody, Christopher Gutteridge, Les Carr, Wendy Hall, Carl Lagoze, and Stevan Harnad. Open citation linking: The way forward. *D-Lib Magazine*, 8(10), October 2002. URL <http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html>. 38
- Steve Hitchcock, Arouna Woukeu, Tim Brody, Les Carr, Wendy Hall, and Stevan Harnad. Evaluating citebase: Key usability results, 2003. URL <http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report-usability.html>. 94

Patrick Hochstenbach, Henry Jerez, and Herbert Van de Sompel. The OAI-PMH static repository and static repository gateway. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 210–217, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-1939-3. URL <https://repository.lanl.gov/retrieve/59/jcdl2003-hochsten-jerez-vandesompel.pdf>. 63

ISI. Master Journal List: Document Solution. Technical report, Thomson Scientific (ISI), 2004. URL <http://scientific.thomson.com/media/presentrep/acropdf/documentsolution.pdf>. “a collection of over 8,700 of the worlds leading peer-reviewed science, technology, social sciences, and arts and humanities journals.”. 28, 54, 117

Evaristo Jiménez-Contreras, Emilio Delgado López-Cózar, Rafael Ruiz-Pérez, and Víctor M. Fernández. Impact-factor rewards affect spanish research. *Nature*, 417:898, 27 June 2002. URL <http://dx.doi.org/10.1038/417898b>. 56

Richard Jones. DSpace vs. ETD-db: Choosing software to manage electronic theses and dissertations. *Ariadne*, (38), January 2004. URL <http://www.ariadne.ac.uk/issue38/jones/>. 71

Simon Kampa. *Who are the experts? e-scholars in the Semantic Web*. Phd thesis, University of Southampton, 2002. URL <http://eprints.ecs.soton.ac.uk/7222/>. 53

Sune Karlsson and Thomas Krichel. Repec and s-wopec: Internet access to electronic preprints in economics, 15 March 1999. URL <http://openlib.org/home/krichel/papers/lindi.html>. 36

LLC Kaufman-Wills Group. The facts about open access: A study of the financial and non-financial effects of alternative business models for scholarly journals. Technical report, ALPSP, Highwire Press and AAAS, 2005. URL <http://www.alpsp.org/publications/pub11.htm>. 23

M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963. 53

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. URL <http://citeseer.ist.psu.edu/kleinberg99authoritative.html>. 105, 161

- Paul Kligfield. Rethinking page charges. *Journal of Electrocardiology*, 38(4): 296–298, 2005. URL <http://dx.doi.org/10.1016/j.jelectrocard.2005.05.004>. Editorial. 15
- Rob Kling and Geoffrey McKim. Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51(14):1306–1320, 2000. URL <http://arxiv.org/abs/cs.CY/9909008>. 15, 16
- Wallace Koehler. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2):174, January 2004. URL <http://informationr.net/ir/9-2/paper174.html>. 44
- M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. Grant, M. Demleitner, E. Henneken, and S. S. Murray. The Effect of Use and Access on Citations. *Information Processing and Management*, 41(6):1395–1402, March 2005. doi: 10.1016/j.ipm.2005.03.010. URL <http://cfa-www.harvard.edu/~kurtz/IPM-abstract.html>. 119, 141
- Michael J. Kurtz. Restrictive access policies cut readership of electronic research journal articles by a factor of two, 2004. URL <http://opcit.eprints.org/feb19oa/kurtz.pdf>. 140, 141
- Martha Kyrillidou and Mark Young. ARL Statistics 2003-04. Technical report, 2004. URL <http://www.arl.org/stats/pubpdf/arlstat04.pdf>. 4, 26
- Carl Lagoze and Herbert Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 54–62, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-345-6. URL <http://doi.acm.org/10.1145/379437.379449>. 39
- Steve Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001. URL <http://citeseer.ist.psu.edu/lawrence01online.html>. 16, 26, 141
- Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999. URL <http://citeseer.ist.psu.edu/article/lawrence99digital.html>. Citeseer <http://citeseer.ist.psu.edu/>. 28, 37

- Xiaoming Liu, Tim Brody, Stevan Harnad, Les Carr, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. A Scalable Architecture for Harvest-Based Digital Libraries: The ODU/Southampton Experiments. *D-Lib Magazine*, 8 (11), November 2002. URL <http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/november02/liu/11liu.html>. vi, 38, 39
- Francisco López-Muñoz, Cecilio Alamo, Gabriel Rubio, Pilar García-García, Belén Martín-Agueda, and Eduardo Cuenca. Bibliometric analysis of biomedical publications on ssri during 1980-2000. *Depression and Anxiety*, 18 (2):95–103, 2003. URL <http://dx.doi.org/10.1002/da.10121>. 53
- Clifford Lynch. Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report*, 226, February 2003. URL <http://www.arl.org/newsltr/226/ir.html>. 22, 73
- John Markwell and David W. Brooks. Broken links: the ephemeral nature of educational WWW hyperlinks. *Journal of Science Education and Technology*, 11(2):105–108, June 2002. URL <http://dx.doi.org/10.1023/A:1014627511641>. 44
- K. Martinez, R. Ong, and J. K. Hart. Glacsweb: a sensor network for hostile environments. In *Proceedings of IEEE 1st International Conference on Sensors and Ad-hoc networks (SECON), 2004*, pages 81–87, London, UK, October 2004. URL <http://www.citebase.org/cgi-bin/citations?id=oai:eprints.soton.ac.uk:15604>. 152
- Philipp Mayr. Constructing experimental indicators for open access documents. *Research Evaluation*, 14, 2006. URL [http://www.ib.hu-berlin.de/~mayr/arbeiten/mayr\\_RE06.pdf](http://www.ib.hu-berlin.de/~mayr/arbeiten/mayr_RE06.pdf). To appear in special issue on 'Web indicators for Innovation Systems'. 118
- M. E. McVeigh. Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns. Technical report, Thomson Scientific, October 2004. URL <http://www.isinet.com/media/presentrep/essayspdf/openaccesscitations2.pdf>. 16
- Michael Miller. Fac Sen discusses journal fees. 6 February 2004. URL [http://daily.stanford.edu/tempo?page=content&id=13027&repository=0001\\_article](http://daily.stanford.edu/tempo?page=content&id=13027&repository=0001_article). 19

- Henk F. Moed. Statistical relationships between downloads and citations at the level of individual documents within a single journal: Book reviews. *J. Am. Soc. Inf. Sci. Technol.*, 56(10):1088–1097, 2005. ISSN 1532-2882. doi: <http://dx.doi.org/10.1002/asi.v56:10>. 118
- Heather Morrison. The elusive art of costing (institutional repositories), 2005. URL [http://poeticeconomics.blogspot.com/2005\\_12\\_01\\_poeticeconomics\\_archive.html](http://poeticeconomics.blogspot.com/2005_12_01_poeticeconomics_archive.html). Blog: The Imaginary Journal of Poetic Economics. 22
- Heather Morrison. The dramatic growth of open access: Implications and opportunities for resource sharing. *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve*, 16(3), 2006. URL <http://ir.lib.sfu.ca/handle/1892/510>. 72
- Michael L. Nelson and B. Danette Allen. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1), January 2002. URL <http://www.dlib.org/dlib/january02/nelson/01nelson.html>. 44
- OAI-PMH. The Open Archives Initiative Protocol for Metadata Harvesting, 2004. URL <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Protocol Document. 63
- Andrew Odlyzko. Tragic loss or good riddance? the impending demise of traditional scholarly journals. *Notices of the American Mathematical Society*, 42:49, January 1995. URL <http://www.ams.org/notices/199501/forum.pdf>. 19
- Andrew Odlyzko. The economics of electronic journals. In R. Ekman and R. Quandt, editors, *Technology and Scholarly Communication*. University of California Press, 1998. URL <http://www.press.umich.edu/jep/04-01/odlyzko.html>. 20
- Open Archives Initiative. The Open Archives Initiative Protocol for Metadata Harvesting Changes from OAI-PMH 1.1 to OAI-PMH 2.0, 2002. URL <http://www.openarchives.org/OAI/2.0/migration.htm>. 62
- Charles Oppenheim. Do Citations Count? Citation Indexing and the Research Assessment Exercise (RAE). *Serials*, 9(2):155–161, 1996. URL <http://dx.doi.org/doi:10.1629/09155>. 29

- Sandra Payette and Carl Lagoze. Flexible and extensible digital object and repository architecture (fedora). In *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 41–59, London, UK, 1998. Springer-Verlag. ISBN 3-540-65101-2. URL <http://www.cs.cornell.edu/lagoze/papers/ECDL98/FEDORA-final.html>. 71
- Sandra Payette and Thornton Staples. The mellon fedora project. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 406–421, London, UK, 2002. Springer-Verlag. ISBN 3-540-44178-6. URL <http://www.fedora.info/documents/ecdl2002final.pdf>. 71
- T. V. Perneger. Relation between online “hit counts” and subsequent citations: prospective study of research papers in the BMJ. *British Medical Journal*, 329: 546–547, 4 September 2004. doi: doi:10.1136/bmj.329.7465.546. URL <http://bmj.bmjournals.com/cgi/content/full/329/7465/546>. 118
- Stephen Pinfield. *International Yearbook of Library and Information Management 2004-2005: Scholarly publishing in an electronic era.*, chapter Self-archiving Publications, pages 118–145. Facet, 2004. URL <http://eprints.nottingham.ac.uk/archive/00000142/01/IYLIM04.PDF>. 20, 72
- William Gray Potter. “Of Making Many Books There is No End”: Bibliometrics and libraries. *The Journal of Academic Librarianship*, 14:238a–238c, 1988. 52
- David Prosser. Science and technology committee report on scientific publications - the uk government’s response. *HEP Libraries Webzine*, (10):5, December 2004. URL <http://library.cern.ch/HEPLW/10/papers/5/>. 25
- S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4(2):131–134, July 1998. URL <http://arXiv.org/cond-mat/9804163>. 51
- Martin Richardson. *Learned Publishing*, 18(3):221–223, July 2005. ‘Personal Views’. 141

- Martin Richardson. Assessing the impact of open access: Preliminary findings from oxford journals. Technical report, Oxford University Press, June 2006. URL [http://www.oxfordjournals.org/news/oa\\_report.pdf](http://www.oxfordjournals.org/news/oa_report.pdf). 24, 143
- Sara Schroter. Importance of free access to research articles on decision to submit to the BMJ: survey of authors. *BMJ*, 332(7538):394–396, 2006. doi: 10.1136/bmj.38705.490961.55. URL <http://bmj.bmjjournals.com/cgi/content/abstract/332/7538/394>. 18
- Per O Seglen. Causal relationship between article citedness and journal impact. *J Am Soc Information Sci*, 45:1–11, 1994. URL [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199401\)45:1%3C1::AID-ASI1%3E3.0.CO;2-Y](http://dx.doi.org/10.1002/(SICI)1097-4571(199401)45:1%3C1::AID-ASI1%3E3.0.CO;2-Y). 56
- Per O Seglen. Why the impact factor of journals should not be used for evaluating research. 314:497, 1997. URL <http://bmj.bmjjournals.com/cgi/content/full/314/7079/497>. 56
- John W T Smith. The importance of access to academic publications for the developing world and the implications of the latest developments in academic publishing, 15-17 September 2004a. URL <http://library.kent.ac.uk/library/papers/jwts/develop.htm>. Presented at the ICCC International Conference on Computer Communication: Core Platform for the Implementation of the Computer Society [Accessed 2006-03-04]. 4
- MacKenzie Smith. DSpace for e-print archives. *High Energy Physics Libraries Webzine*, CERN, page 7pp, September 2004b. URL <http://eprints.rclis.org/archive/00001503/>. 71
- Mike Gardner Stephen Pinfield and John MacColl. Setting up an institutional e-print archive. *Ariadne*, (31), 2002. URL <http://www.ariadne.ac.uk/issue31/eprint-archives/>. 73
- Peter Suber. NIH public-access policy: Frequently asked questions, February 6 2006. URL <http://www.earlham.edu/~peters/fos/nihfaq.htm>. 25
- Danny Sullivan. Google Scholar offers access to academic information. 18 November 2004. URL <http://searchenginewatch.com/searchday/article.php/3437471>. [Accessed 2006-03-07]. 27



- Alma Swan and Sheridan Brown. Open access self-archiving: An author study. Technical report, Joint Information Systems Committee (JISC), UK FE and HE funding councils, 2005. URL <http://cogprints.org/4385/>. 18, 26
- Janice Hopkins Tanne. Researchers funded by NIH are failing to make data available. *BMJ*, 332(7543):684–b–, 2006. doi: 10.1136/bmj.332.7543.684-b. URL <http://bmj.bmjournals.com>. 25
- Timothy W. Cole Thomas G. Habing and William H. Mischo. Developing a technical registry of OAI data providers, September 2004. URL <http://www.ecdl2004.org/presentations/session-10a-paper-2/>. Presented at European Conference on Digital Libraries 2004, University of Bath, 12-17 September 2004. 75
- Thomson ISI. The Thomson Scientific journal selection process, January 2004. URL <http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/>. Accessed 2006-03-13. 55
- UK Parliament Science and Technology Committee Publications. Scientific Publications: Free for all?, 2004. URL <http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm>. 25
- UK RAE. RAE 2008 Panel criteria and working methods. Technical report, UK Research Assessment Exercise, January 2006. URL <http://www.rae.ac.uk/pubs/2006/01/>. 29, 30
- US House Appropriations Committee. House report 108-636, 2004. URL [http://thomas.loc.gov/cgi-bin/cpquery/?&db\\_id=cp108&r\\_n=hr636.108&sel=TOC\\_338641](http://thomas.loc.gov/cgi-bin/cpquery/?&db_id=cp108&r_n=hr636.108&sel=TOC_338641). US House Appropriations Committee, Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriation Bill, 2005. 25
- Herbert Van de Sompel and Oren Beit-Arie. Open linking in the scholarly information environment using the openurl framework. *D-Lib Magazine*, 7(3), March 2001. URL <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>. 45
- Herbert Van de Sompel and Carl Lagoze. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2), 2000. URL <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>. 2, 27, 39

Herbert Van de Sompel and Carl Lagoze. Notes from the interoperability front: A progress report on the Open Archives Initiative. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 144–157, London, UK, 2002. Springer-Verlag. ISBN 3-540-44178-6. 39

Lilian Van de Vaart. DARE, the voyage begun. Technical report, Eleftheria: Knowledge Management, 2004. URL <http://eprints.rclis.org/archive/00003569/>. 89

Annemiek Van der Kuil and Martin Feijen. The dawning of the dutch network of digital academic repositories (DARE): A shared experience. *ARIADNE*, (41), October 2004. URL <http://www.ariadne.ac.uk/issue41/vanderkuil/>. 88, 89

Virginia Tech. ETD-db home page, 2006. URL <http://scholar.lib.vt.edu/ETD-db/>. 71

Jenny Walker. Openurl and SFX linking. *Serials Librarian*, 45(3):87–100, 2003. URL [http://www.exlibrisgroup.com/resources/sfx/OpenUR\\_SFX\\_for\\_Serials\\_Librarian\\_Nov\\_2003.pdf](http://www.exlibrisgroup.com/resources/sfx/OpenUR_SFX_for_Serials_Librarian_Nov_2003.pdf). 45

Jewel Ward. A quantitative analysis of unqualified Dublin Core Metadata Element Set usage within data providers registered with the Open Archives Initiative. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on*, pages 315–317, May 27–31 2003. ISBN 0-7695-1939-3. URL <http://www.foar.net/research/mp/wardj-quantitative2.pdf>. 70

Simeon Warner. The OAI Data-Provider Registration and Validation Service, 2005. URL <http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0506010>. 65

Wellcome Trust. Wellcome trust position statement in support of open and unrestricted access to published research, February 9 2006. URL [http://www.wellcome.ac.uk/doc\\_WTD002766.html](http://www.wellcome.ac.uk/doc_WTD002766.html). 25

John White. ACM digital library enhancements. *Commun. ACM*, 42(4):30–31, 1999. ISSN 0001-0782. URL <http://doi.acm.org/10.1145/299157.299519>. 34

- John White. ACM opens portal. *Commun. ACM*, 44(7):14–ff, 2001. ISSN 0001-0782. URL <http://doi.acm.org/10.1145/379300.379304>. 34
- Sonya White and Claire Creaser. Scholarly journal prices: Selected trends and comparisons. Technical report, Loughborough University, UK, October 2004. URL <http://www.lboro.ac.uk/departments/dis/lisu/pages/publications/oup.html>. 19
- Rodney Yancey. Fifty years of citation indexing and analysis. *KnowledgeLink Newsletter*, August/September 2005. URL <http://scientific.thomson.com/news/newsletter/2005-08/8289803/>. 32
- George K. Zipf. Addison-Wesley, Cambridge MA, 1949. 51

# Index

- ACM Digital Library, 34
- archive
  - definition, 9
  - institutional repositories, 21
- arXiv, 118
- Bibliographic Co-Citation, 53
- Bibliographic Coupling, 53
- Bradford's Law, 53
- citation
  - definition, 9
- Citeseer, 37
- Co-Citation, *see* Bibliographic Co-Citation
- copyright
  - self-archiving, 21
- CrossRef, 47
- Digital Object Identifier (DOI), 47
- Dublin Core, 41
- e-print
  - definition, 9
- e-print archive, *see* archive
- Eugene Garfield, 54
- Impact Factor, *see* Journal Impact Factor
- institutional archives, *see* institutional repositories
- institutional repositories, 21
- Journal Impact Factor, 55
- Lotka's Law, 51
- metrics
  - research evaluation, 29
- OAI-PMH, 39
- open access
  - definition, 9
- Open Archives Initiative, The, 39
- Open Citation Project, The, 38
- OpenURL, 45
- paper
  - definition, 9
- reference
  - definition, 8
- RePEc, 36
- research assessment exercise (RAE), 29
- ResearchIndex, 37
- Science Citation Index
  - history, 54
  - Web of Science, 32
- ScienceDirect, 34
- self-archiving, 20
- serials crisis, 19
- Web of Science, 32
- Zipf's Law, 51